(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
24 December 2003 (24.12.2003)

PCT

(10) International Publication Number
WO 03/106635 A2

(51) International Patent Classification[7]: C12N

(21) International Application Number: PCT/US03/18714

(22) International Filing Date: 13 June 2003 (13.06.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/387,887 13 June 2002 (13.06.2002) US
60/387,910 13 June 2002 (13.06.2002) US

(71) Applicant *(for all designated States except US)*: **REGULOME CORPORATION** [US/US]; Canal View Building, 551 N. 34th Street, Seattle, WA 98103 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **STAMATOYANNOPOULOS, John, A.** [US/US]; 15 Milford Street, #1, Boston, MA 02118 (US). **MCARTHUR, Michael** [GB/GB]; Mill House, Mill Lane, Rockland NR17 1XR (GB). **SABO, Peter, J.** [US/US]; Canal View Building, 551 N. 34th Street, Seattle, WA 98103 (US).

(74) Agents: **ANTLER, Adriane, M.** et al.; Pennie & Edmonds LLP, 1155 Avenue of the Americas, New York, NY 10036 (US).

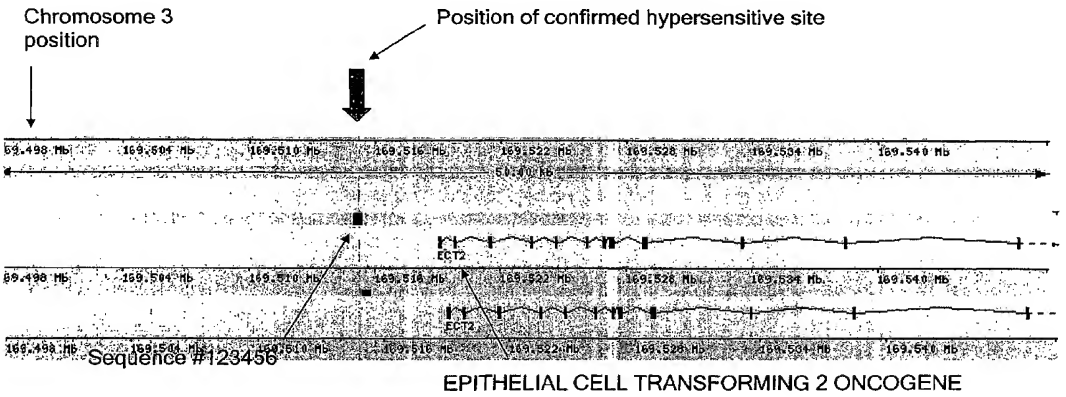
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: FUNCTIONAL SITES



(57) Abstract: Functional site sequences, their genomic positions, and methods of their use are described. The sequences are individually useful for their abilities to regulate gene expression. Sets and arrays of sequences are particularly useful for the identification of cells and tissues, the detection and diagnosis of disease, and the discovery of medicinal agents, and gene regulatory mechanisms, including those associated with disease. The sequences and their locations also are highly valuable when used by computer programs for comparing known genetic sequences by a large variety of computer manipulations.

WO 03/106635 A2

FUNCTIONAL SITES

This application claims benefit under 35 U.S.C. § 119(e) of U.S. provisional application nos. 60/387,910 and 60/387,887, both filed on June 13, 2002, each of which is incorporated by reference herein in its entirety.

FIELD OF THE INVENTION

The invention relates generally to functional sites identified within the genome and their use in the diagnosis and treatment of diseases, including immuno-compromised disorders, neurological disorders, genetic disorders, cancers and infectious diseases.

BACKGROUND OF THE INVENTION

Despite great advances in sequencing the human genome, the regulation of human genes is poorly understood. It has been known for many years that most DNA does not encode structural genes and most DNA transcription creates RNA that never leaves the nucleus. However, relatively little is known about the function of non-coding genomic DNA. Although a large number of such DNA sequence appears to be involved in gene regulation, details of the regulatory DNA sequences and their locations remain mostly unknown. This unfortunate situation may be a natural result of an emphasis on protein encoding genes and the search for sequences defined by open coding regions.

GENE REGULATION

Understanding the human genome will require comprehensive identification of DNA elements that are functional *in vivo*. A major class of such sequences are those which have a role in regulating genomic activity.

Regulatory factors interact with chromatin in a site-specific fashion to bring the genome to life. All genes are controlled at multiple levels through the interaction of regulatory factors with gene-proximal or, in some cases, distant *cis*-

regulatory sites. The nucleoprotein complexes formed by such interactions may be tissue or developmental stage-specific, or they may be constitutive, depending on the regulatory requirements of their cognate gene. While our knowledge of the patterns of gene expression in diverse tissues and under a wide-ranging set of conditions has grown substantially in recent years, this growth has not been paralleled by a comparable increase in our knowledge of regulatory factors that control specific genes affecting specific cellular or disease processes.

CONTROL OF GENE EXPRESSION *IN VIVO*

The basic chromatin fiber consists of an array of nucleosomes, each packaging around 200 base pairs of DNA; 146 is wound around the histone octamer, with the remainder forming a link to the next nucleosome. In eukaryotic cells, all genomic DNA in the nucleus is packaged into chromatin, the architecture of which plays a central role in regulating gene expression (for reviews see Felsenfeld and Groudine 2003; Felsenfeld, 1992; Brownell and Allis, 1996; Kingston *et al.*, 1996; Tsukiyama and Wu, 1997; Wolffe *et al.*, 1997; Kadonaga, 1998; Struhl, 1998). At a global level, this packaging serves two purposes: (i) it is physically necessary to condense the mass of sequence information into a well-ordered regular structure that can be contained within the nucleus; and (ii) it imparts a level of site-specific 'epigenomic' information (Felsenfeld, 1992), for example discriminating between sequences which are never to be transcribed and are stored in highly condensed heterochromatin, and those sequences which are actively transcribed and are maintained in a more accessible chromatin state.

Gene expression is regulated by several different classes of *cis*-regulatory DNA sequences including enhancers, silencers, insulators, and core promoters (Felsenfeld and Groudine, 2003; Butler and Kadonga, 2002; Gill, 2001). The core promoter is the site of formation of the RNA pol II transcription complex. Enhancers and silencers act over distances of several kilobases (or more) to potentiate or silence pol II function. Insulator sequences prevent enhancers and silencers targeted to one gene from inappropriately regulating a neighboring gene.

Larger more complex elements comprising multiple enhancer and/or silencers have come to light which coordinate the activity of linked genes over large chromosomal domains ('Locus Control Regions' or 'Domain Control Regions') (reviewed in Li *et al.*, 2002a, Hardison, 2001). Activation of *cis*-regulatory elements in the context of chromatin requires the cooperative binding of regulatory factors (Felsenfeld, 1996). This active state is most commonly addressed by measuring the sensitivity of the underlying DNA sequences to digestion with nucleases (e.g., DNaseI) in the context of chromatin (Weintraub and Groudine, 1976; Elgin, 1981). Multiprotein complexes exist in cells that allow specific destabilization of nucleosomes at promoters, facilitating the binding of sequence-specific factors and the general transcriptional machinery (Kingston *et al.*, 1996; Svaren, 1996; Tsukiyama and Wu, 1997). Posttranscriptional modifications of chromatin components, particularly histone acetylation, play important roles in regulating chromatin structure and gene activity (Brownell and Allis, 1996; Grunstein, 1997; Wolffe *et al.*, 1997; Kadonaga, 1998; Struhl, 1998).

Activation of tissue-specific genes during development and differentiation occurs first at the level of chromatin accessibility and results in the formation of transcriptionally-competent genetic loci characterized by increased sensitivity (relative to inactive loci) to digestion with DnaseI (Groudine *et al.,* 1983; Tuan *et al.,* 1985; Forrester *et al.*, 1986). Loci in an accessible chromatin configuration can subsequently respond to acutely activating signals, often conveyed by non-tissue-specific transcriptional factors that can gain access to the open locus and recruit or activate the basal transcriptional machinery.

The initial observation that active genes reside within domains of generally increased sensitivity to nucleases was made nearly 30 years ago (Weintraub and Groudine, 1976). Since this time, such data have been accumulated for a number of human gene loci (Pullner *et al.,* 1996) and those in other vertebrates (Koropatnick and Duereksen, 1987 Stratling *et al.*, 1986). The chromatin domain phenomenon is particularly striking in *Drosophila*, where distinct

transitions between DNase-sensitive and DNase-resistant chromatin can be documented (Farkas *et al.*, 2000).

DNase I Hypersensitive Sites Identify Active Regulatory Regions *in vivo*. Focal alterations in chromatin structure are the hallmark of active regulatory sequences in eukaryotic genomes. The literature connecting DNaseI-hypersensitive sites with genomic regulatory elements is extensive. DNase hypersensitivity studies have been employed to delineate the transcriptional regulatory elements of over 100 human gene loci . Typically, between 1 and 5 hypersensitive sites have been visualized for each of these loci. However, only a fraction of these have been precisely localized at the sequence level.

Nuclease hypersensitivity studies represent a powerful, *in vivo* approach to detection and analysis of biologically active sequences. A critical defining feature of HSs is that the function of the DNA sequence component – i.e. its complex-forming activity – is intrinsic. The principal evidence for this is the fact that these sequences can be excised and inserted into other positions in the genome, where they exhibit the same functional chromatin activities. Substantial experimental experience from model systems has revealed that HSs can form when included in either constructs used to create stably transfected cell lines (Fraser *et al.*, 1990) or transgenic animals (Lowrey *et al.*, 1992; Levy-Wilson *et al.*, 2000).

An important finding has been that HS sequences are rendered functional only upon assembly into nuclear genomic chromatin. These DNA sequences are thought to potentiate formation of a nucleoprotein complex in a manner that dramatically increases its probability of activation vs. neighboring DNA regions. They are hypothesized to adopt a particular topological confirmation, which lowers the free energy for coalescence of a limited set of proteins, some in contact with DNA, and some in contact only with another protein in the complex. This results in the formation of a nucleoprotein complex which is precisely correlated with a particular sequence. The formation of this complex takes place in

an 'all-or-none' fashion (e.g., Felsenfeld *et al.*, 1996; Boyes & Felsenfeld, 1996). The stochasticity of nucleoprotein complex formation can be manipulated through the introduction of point mutations or small deletions or insertions in critical DNA binding bases or in juxtaposed sequences that affect overall stability (e.g., Stamatoyannopoulos *et al.*, 1995).

Study of DNaseI hypersensitive sites has shown that they are commonly associated with functional activities important in regulating genome biology. These activities (summarized in Table 2 below) include control of transcription, replication and architecture: transcriptional promoters (Levy-Wilson 2000), enhancers (Furbass 2001), Matrix Attachment Regions (MARs; van Drunen et al., 1999), chromatin insulators (Li et al., 2002b), transcriptional silencers (Youn et al. 2002) and origins of replication (Aladjem et al., 1998).

Nuclease hypersensitive sites are biologically bounded by (1) the positions of flanking nucleosomes and (2) limits on the area of DNA over which thermodynamically stable nucleoprotein complexes may form. The extent of the regulatory domain is contained within the inter-nucleosomal interval, approximately 150-250bp. This interval corresponds to the size of sequence that is needed to place a canonical nucleosome and it has been a common assumption that HSs represent a break in the nucleosomal array that constitutes the vast majority of chromatin.

A core domain can be identified which is restricted to a region of approximately 80-120 base pairs in length, over which DNA-protein interactions take place (e.g., Lowrey *et al.*, 1992). Cooperative binding of transcription factors to such core regions is sufficient to exclude a nucleosome *in vitro* (Adams and Workman, 1995) and this has been proposed as a common mechanism for how these sites may form *in vivo*. Nucleosomal mapping experiments have shown that HSs such as the *Drosophila hsp26* promoter (Lu *et al.*, 1995) and the human β-globin HS2 (Kim and Murray, 2001) are non-nucleosomal. It is thought that most HSs are non-nucleosomal in nature (Boyes and Felsenfeld, 1996; Wallrath *et al.*,

1994). These conclusions are well-supported in the literature (e.g., *ibid* and Struhl, 2001). However several HSs are known to still have histone proteins and transcription factors, suggesting that HSs may exist in conjunction with a modified or partial nucleosome.

Flanking sequences surrounding the core region appear to modulate the activity of this core region, though this effect tapers off sharply. The boundaries of the sequences needed for hypersensitivity can be defined functionally by performing deletion analyses followed by stable transfection of cells (Philipsen *et al.*, 1993) or transgenic studies (Lowrey *et al*, 1992; Zhou *et al.*, 1995). These approaches define the minimum extent of sequence required to retain the biological function associated with the HS under examination.

It is observable that many hypersensitive sites occur within broader domains of increased DNase sensitivity and therefore appear to be components of higher-order chromatin structures. It is further observable that, based on published data, such sites appear to harbor increased biological significance and are perhaps the most important functionally. Several investigators have observed that the regions flanking the hypersensitive foci of active elements exhibit an increased level of sensitivity to nuclease digestion compared with the increased general sensitivity of an active locus. This phenomenon has been referred to as 'intermediate sensitivity' (Kunnath and Locker, 1985).

In summary, active genes are embedded within broad regions of increased chromatin accessibility punctuated by foci of hyper-accessibility that coincide with active regulatory sequences. All genomic sequences to which regulatory factors are complexed in vivo will be expected to produce focal alterations in chromatin structure that can be detected via nuclease hypersensitivity studies

GENETICS AND GENE EXPRESSION

Overview. Most common diseases are polygenic and due to quantitative variation in particular phenotypic traits. The principal mechanism by

which quantitative variation is affected is through the regulation of gene expression. Therefore, it is expected that a substantial proportion of functional genetic mutations that cause or modulate common diseases will be found in *cis*-regulatory sequences (Rockman 2002).

The functional impact of sequence variants in coding sequences is buffered by (i) the degeneracy of genetic code; (ii) the similarity of behavior between certain classes of amino acids (leading to 'conservative' vs. 'radical' substitutions); and (iii) the fact that important functional regions occupy only a small percentage of the protein sequence.

In contrast, a single nucleotide lesion within a sequence comprising docking site for a DNA-binding factor can fully abolish the capacity of this site to serve as a regulatory region, with consequent deleterious effects on gene function and phenotype. Several such examples are known (Knight 1999; Prokunina 2002; Wu 2003; Zwarts 2002) and more are coming to light with increasing frequency.

The number of *cis*-regulatory sites in the human genome is unknown. Based on the observation that genes – particularly those which are highly regulated during development, differentiation, or in response to pharmaceuticals – have multiple *cis* elements, the total number of such sites in the genome is expected to be a multiple of the number of genes.

Gene regulation and quantitative traits

Genetic analyses look for differences in gene sequence that could explain variation in physical traits. Gene-expression studies provide a snapshot of active genes. These approaches can be combined to great effect.

Complex traits — physical or behavioural characteristics of an organism – are generally dictated by combinations of more than one gene and the environment. *Combining studies of functional sites with gene expression, complex traits may be studied at a greater scale and depth than is possible using either technique alone.*

Genome-wide genetic analyses of gene-expression data was introduced by Jansen and Nap, and Brem *et al.* were the first to apply this approach, in a study of budding yeast. Variations in DNA sequence across the genomes of a population are analyzed to identify their origin (that is, which one of the two progenitor strains). At the same time the population is studied to find out which genes are being expressed in different individuals, and to what degree. The expression level of each gene is then treated as a "quantitative trait".

Quantitative traits are typically determined by more than one gene and show a graded variation across a population, such that the variation can only be measured quantitatively. Height, weight, and blood pressure are typical examples. Since variations in regulatory sequences are inherited, gene-expression levels can also be considered to be a quantitative trait. It is therefore expected that genetic changes that control gene expression will reside primarily in the same chromosomal region as the genes that are controlled.

Statistical analyses may be carried out to correlate DNA variations with gene-expression levels. A statistically significant correlation suggests that the gene (or genes) in the chromosomal region where the sequence variation occurs may account for some of the variation in gene expression. As the process is carried out for the entire genome, the results might highlight previously unknown gene–gene interactions, identify biochemical pathways and enable genetically like individuals to be grouped together. This last point in particular could be relevant to 'personalized' medicine — the development of drugs that are tailor-made for specific groups of patients.

Certain regulatory elements may affect the expression levels of several genes within a genomic domain. Many examples of this have been documented, the most extensively studied of which is the beta-globin locus on chromosome 11. Such elements may control how the expression levels of different genes in a given pathway are correlated.

An additional degree of sophistication can be introduced to genetic gene expression analyses by including a particular trait — a disease, for instance. Gene-expression data might help to define such a trait more accurately, generating genetically more homogeneous groups of individuals that have that characteristic. Genetic analysis of these subgroups would then permit identification of chromosomal regions that influence a quantitative trait.

The combination of gene-expression and genetic data could also have another use: it might help to identify candidate genes that affect a given trait. This could be achieved by looking for overlaps between differentially expressed genes and variation in functional sites, and by looking at functional sites that are found in a common chromosomal region. The approach may be generalized to almost any organism in which both gene-expression profiling and genome-wide genetic analysis can be done efficiently.

TRANSCRIPTIONAL REGULATION IN EVOLUTION

The role of variation in functional sites in shaping evolution of complex traits is widely recognized. This has been comprehensively reviewed by Wray et al (2003) from which the following summary can be made:

Several recent reviews have argued that changes in transcriptional regulation comprise a major component of the genetic basis for phenotypic evolution (Doebley and Lukens 1998; Carroll 2000; Stern 2000; Tautz 2000; Theissen et al. 2000; Purugganan 2000; Wray and Lowe 2000; Carroll et al. 2001; Davidson 2001; Wilkins 2002).

Mutations in transcriptional regulation influence phenotype. Transcriptional regulation is an integral component of the way genotype is converted into phenotype. Many mutants that have emerged from genetic screens for developmentally important genes involve defects in transcriptional regulation (Wilkins 1993, 2002; Gilbert 2000). The four-winged fly that results from certain mutations in *Ubx* in *Drosophila* is perhaps the most famous: some mutations located in regulatory sequences affect the transcription profile, while others

locating in exons alter the function of the protein in regulating the transcription of other genes (Bender et al. 1983; Simon et al. 1990). The phenotypic consequences of some *Ubx* functional site mutations are so distinct that they were originally thought to represent separate genes (Lewis 1978). Numerous studies have documented correlations between gene expression and anatomy. (1) *Induced mutations*. The phenotypes of some induced mutations mimic natural differences between species. Examples include homeotic mutations in d*rosophila melanogaster* that mimic segment and appendage number and identity characteristic of other insects (Raff and Kaufman 1983; Carroll 1995), mutations in *Arabidopsis thaliana* and *Antirrhinum majus* that mimic the floral anatomy of other angiosperms (Lawton- Rauh et al. 2000), and mutations in *Caenorhabditis elegans* that mimic the tail anatomy of other nematodes (Fitch 1997). Because most of these induced mutations generally do not replicate the genetic basis for natural phenotypic differences (Carroll 1995; Budd 1999), however, convincing evidence of the evolutionary significance of changes in transcriptional regulation must come from natural cases. (2) *Comparisons of expression*. In many cases, a gene required for the development of a trait in one species shows a difference in expression in other species that correlates with a difference in that trait (e.g., Burke et al. 1995; Brakefield et al. 1996; Dudareva et al. 1996; Sinha and Kellogg 1996; Averof and Patel 1997; Stockhaus et al. 1997; Abzhanov and Kaufman 2000; Kopp et al. 2000; Yamamoto and Jeffery 2000; Beldade et al. 2002; Bharathan et al. 2002; Hariri et al. 2002). A causal relationship is plausible but not proven in these cases, since comparisons of gene expression cannot by themselves demonstrate that a change in transcriptional regulation is the genetic basis for a phenotypic difference. (3) *Quantitative genetics*. Anatomical changes that accompanied the domestication of maize from teosinte are due in part to changes within the inferred functional site region of a single gene encoding the transcription factor teosinte-branched (Wang et al. 1999). Although this is a case of artificial selection, it involved natural (rather than induced) genetic variation. Some differences in bristle

patterns among *Drosophila* species are attributable to changes in functional site sequences (Stern 1998; Skaer and Simpson 2000; Sucena and Stern 2000). In other cases, genetic variation in gene expression levels shows strong associations with specific organismal phenotypes (Gerber et al. 2000; Karp et al. 2000; Beldade et al. 2002). Unfortunately, quantitative genetics generally lacks the resolution to identify precise sequence differences that are responsible for particular phenotypes, due to the confounding effects of linkage disequilibrium. When combined with experimental tests or case associations, however, specific sequence variants can be identified (Cooper 1999). Using this approach, more than 160 segregating functional site variants that influence transcription have been identified in humans (Cooper 1999; Rockman and Wray 2002) and several have been identified in *Drosophila melanogaster* (e.g., Robin et al. 2002).

Natural populations harbor considerable functional variation in gene expression. Many examples of variation in gene expression are known from natural populations. (1) *Spatial extent of expression.* In rainbow trout, an allele of PGM1 conferring expression in the liver is associated with faster pre-hatching growth (Allendorf 1982, 1983). The spatial expression of amylase in the midgut varies within both *Drosophila melanogaster* and *D. pseudoobscura*; the genetic basis in both cases is *trans* and responds to artificial selection in *D. pseudoobscura* (Abraham and Doane 1978; Powell and Licthenfels 1979; Powell 1979). The spatial extent of expression of the transcription factor Distal-less within the wing of the butterfly *Bicyclus anynana* varies in correlation with wing color pattern, and also responds to artificial selection (Beldade et al. 2002). (2) *Level of expression.* Intraspecific differences in expression have been noted for GPDH in both larvae and adults of *D. melanogaster* (Laurie-Ahlberg and Bewley 1983); □-*glucuronidase* in *Mus domesticus* (Pfister et al. 1982; Bush and Paigen 1992); *Cyp6g1*, a *cytochrome P450* family gene, in *D. melanogaster* (Daborn et al. 2002); and *prolactin* in the teleost *Oreochromis niloticus* (Streelman and Kocher 2002). In all four cases, most or all of the polymorphisms described are in *cis*. Many

additional examples are known from humans, where nearly 2/3 of the known functional polymorphisms in *cis*-regulatory sequences have a >2-fold impact on transcription rates (Rockman and Wray 2002). (3) *Inducibility of expression*. Inducibility of amylase expression in response to a starch diet varies within *D. melanogaster* and responds to artificial selection (Matsuo and Yamazaki 1984; Klarenberg et al. 1987); expression of □-*glucuronidase* in response to androgen varies within *Mus domesticus* (Bush and Paigen 1992); and three different mobile element insertions into the promoter of *hsp70* reduce transcription in response to thermal stress in *D. melanogaster* populations (Lerman et al. 2003). Several other examples of variation in inducibility are known from humans (Rockman and Wray 2002). In the human and *hsp70* cases, the genetic basis is known to reside in *cis*. Additional studies have estimated the extent of heritable genetic variation in gene expression within populations. (1) *Protein-based surveys*. Several studies have measured levels of variation in gene expression from 1- or 2-dimensional protein gels in a variety of organisms: *Zea mays* (Burstin et al. 1994; Damerval et al. 1994; de Vienne et al. 2001), *Pinus pinaster* (Costa and Plomion 1999), *Glycine max* (Gerber et al. 2000), *Mus musculus* (Klose et al. 2002), and *Homo sapiens* (Enard et al. 2002a). Studies with the first three organisms documented that protein abundance has a strong genetic component, while all of these studies found that populations contain considerable variation in expression level for most of the proteins surveyed. In *D. melanogaster*, chromosome substitution lines show substantial levels of variation in gene expression as measured by enzyme activities (Laurie-Ahlberg et al. 1980; Wilton et al. 1982; Clark 1990). Although protein abundance and enzyme activity are indirect indices of transcription, these results suggest considerable genetic variation for gene expression in general. (2) *mRNA-based surveys*. More direct estimates of variation in transcription come from microarray analyses that survey thousands of loci. Studies in mice (Karp et al. 2000; Schadt et al. 2003), humans (Schadt et al. 2003), the teleost *Fundulus heteroclitus* (Oleksiak et al. 2002), *D. melanogaster* (Jin et al. 2001; Rifkin et al.

2003), *Zea mays* (Schadt et al. 2003), and *Saccharomyces cerevisiae* (Cavalieri et al. 2000; Brem et al. 2002) all indicate that genetic variation in transcript abundance is pervasive within populations. Much of this variation may be heritable. Schadt et al. (2003) found that 33% of the 23,574 loci surveyed from a cross of two inbred strains of mice showed a genetic component for expression differences within the liver, 29% of the 2,726 loci surveyed from 56 humans belonging to four families showed a heritable difference in expression within lymphoblasts, and 18,805 genes consistently differed in transcription within ear leaf tissue among progeny from a cross of two maize strains. What proportion of the genetic basis for this variation resides in the functional sites of the genes showing transcriptional variation (*cis*) or in the sequences or expression profiles of their upstream regulators (*trans*) has been examined in a few cases. QTL underlying variation in expression of at least 32% of 570 variably-expressed transcripts in yeast mapped in *cis* (Brem et al. 2002), while the comparable fraction of genes with *cis*-acting QTL in mouse liver is even higher (Schadt et al. 2003). RT-PCR offers more reliable quantitation than microarrays as well as the ability to compare transcription rates among alleles directly. In a preliminary survey of 69 loci in four inbred lines of *Mus musculus*, Cowles et al. (2002) found quantitative and tissue-specific variation among alleles at 4 loci. Using a similar approach, Yan et al. (2002) found evidence of variation in gene expression at 6 of 13 loci examined in humans. Taken together, microarray and RT-PCR surveys of mRNA levels provide solid evidence of abundant genetic variation in transcriptional regulation in diverse species and suggest that much of this variation resides in *cis* regulatory sequences. (3) *Detailed analyses of functional site function*. The most extensive direct evidence of functional variation in functional site sequences currently comes from humans, where many specific polymorphisms have been identified through direct functional studies (Cooper 1999). Although the human genome is not particularly polymorphic, a typical individual is estimated to be heterozygous for a functional functional site polymorphism at ~40% of all loci (Rockman and Wray 2002).

Comparable data do not yet exist for other species, but RT-PCR surveys (Cowles et al. 2002; Yan et al. 2002) provide a rapid means of estimating heterozygosity that affects transcription at many loci.

Natural selection operates on allelic variation in functional sites. Evidence for natural selection on eukaryotic functional site alleles comes from a variety of sources. (1) *Human populations*. Functional site polymorphisms at numerous loci in humans have functional consequences, influencing diverse aspects of physiology, behavior, anatomy, and life history (Cooper 1999; Rockman and Wray 2002). Some of these functional site alleles have likely fitness consequences. (2) *Wild populations*. A latitudinal cline of LDH functional site allele frequencies in the teleost *Fundulus heteroclitus* is probably maintained by temperature differences (Crawford et al. 1999; Segal et al. 1999). Two other cases, mentioned earlier, are known from *D. melanogaster*: functional site alleles segregating at both *Cyp6G1* and *hsp70* appear to be under selection in wild populations (Daborn et al. 2002; Lerman et al. 2003). (3) *Artificial selection and experimental evolution*. Domestication of maize involved selection on the inferred regulatory region of *tb* locus (Wang et al. 1999). Studies with yeast point to regulation of transcription as a critical component of adaptive change. Adaptation of *Saccharomyces cerevisiae* to glucose limitation was accompanied by two-fold or greater changes in the abundance of transcripts from nearly 10% of all genes, consistently across replicates (Ferea et al. 1999). The evolution of drug resistance in experimental populations of *Candida albicans* correlated with overexpression of the four known resistance genes (Cowen et al. 2000). (4) S*equence comparisons.* More extensive, but less direct, evidence that natural selection acts on functional sites comes from cases of apparent evolutionary conservation of *cis*-regulatory sequences among distantly related species. Consistent under-representation of specific sequence motifs provides evidence for genome wide selection to remove spurious transcription initiation sequences in a broad diversity of prokaryotes (Hahn et al. 2003). Several

examples of natural selection operating on transcriptional regulation involve pathogen-host interactions. For instance, some promoter alleles in *Mycobacterium tuberculanum* and hepatitis B alter transcription to the pathogen's benefit and may be under positive selection (Buckwold et al. 1997; Rinder et al. 1998; Lee et al. 2000; Kajiya et al. 2001). The origin and subsequent fixation of these mutations in separate host individuals demonstrates the ability of positive selection to operate in a predictable way on genetic variation within a functional site. Specific variants within the HIV promoter, including gains of binding sites for host NF-kB and USF transcription factors, as well as functional modifications in the basal promoter, cause differences in the level of viral transcription (Montano et al. 1997; Jeeninga et al. 2000). The E subtype of HIV has significantly increased transcription rates and has gone to near fixation locally in southern Africa; it is associated with increased levels of secondary infections and may be under positive selection to the pathogen's advantage (Montano et al. 2000; Hunt et al. 2001).

Conversely, human populations harbor functional site variants that influence susceptibility to pathogens or disease progression following infection. Because human generation times are much longer than those of pathogens, signatures of selection are more difficult to detect. Nonetheless, functional site alleles at *TNF* □, *IL-4*, *IL-10*, *FY*, *CCR5*, and *TGF* □influence mortality from a variety of viral, bacterial, and protoctistan pathogens and are likely to be under selection (Tournamille et al. 1995; Hamblin and Di Rienzo 2000; Thursz 2001; Bamshad et al. 2002; Meyer et al. 2002; Shin et al. 2000; Vidigal et al. 2002; Nakayama et al. 2002). Some functional site alleles confer protection from one pathogen while increasing susceptibility to another (e.g., *TNF* □–380*A*: Meyer et al. 2002), raising the possibility of balanced polymorphisms.

Changes in functional sites arise from a variety of mutations. Mutations affecting transcription fall into several distinct classes. (1) *Small-scale, local mutations can modify, eliminate, and generate binding sites and alter their spacing*. Functional site function can be directly altered by the most abundant

kinds of mutations: single base substitutions, small indels, and changes in repeat number (e.g., Segal et al. 1999; Gonzalez et al. 1995; Shashikant et al. 1998; Takahashi et al. 2001; Rockman and Wray 2002; Streelman and Kocher 2002). Point mutations can modulate or eliminate transcription factor binding, generate binding sites *de novo* , or result in binding by a different transcription factor ("transcription factor switching": Rockman and Wray 2002). Insertions and deletions can change spacing between binding sites, as well as eliminate binding sites or generate new ones (Ludwig and Kreitman 1995; Belting et al. 1998). Changes in microsatellite structure can affect spacing between binding sites and alter the number of binding sites, sometimes with functional consequences (Trefilov et al. 2000; Rockman and Wray 2002; Streelman and Kocher 2002). (2) *New regulatory sequences can be inserted into functional sites through transposition*. This phenomenon has been reviewed extensively (Kidwell and Lisch 1997; Britten 1997; Brosius 1999). For instance: B2 SINEs in *Mus musculus* contain sequences capable of acting as basal promoters (Ferrigno et al. 2001) and some Alu elements in humans contain binding sites for nuclear hormone receptors and exert an influence on transcription (Babich et al. 1999). (3) *Retroposition may assemble new functional sites*. Retroposition can create novel genes that are subsequently expressed (e.g., *jingwei* and *sphinx*: Long et al. 1999; Wang et al. 2002). This process occurs at appreciable frequencies within the genus *Drosophila* (Bétran et al.2002). The molecular mechanisms underlying retroposition preclude transfer of the basal promoter and virtually all *cis*-regulatory sequences (the exception being those within exons). Because no gene can function without transcriptional regulatory sequences, it seems likely that novel genes that arise through retroposition either fortuitously insert near existing *cis*-regulatory sequences and come under their regulation without disrupting existing regulatory fuctions, or persist long enough that novel *cis*-regulatory sequences arise through transposition, recombination, or small scale local mutations. Remarkably, novel genes that arise through retroposition are often expressed in

tissue-specific patterns similar to those of a parent locus (Bétran et al. 2002). (4) *Gene duplications may fragment or recombine functional site sequences.* Although analyses of gene duplication typically focus on coding sequences, the associated functional sites are clearly also important for gene function. If the breakpoints do not include *cis*-regulatory sequences, the duplicated copy is likely to be transcriptionally inert in its new location and become a pseudogene even before it accumulates stop codons or frameshifts. If only part of the functional site is duplicated, the transcription profile of the new copy may differ from the original (e.g., *nNOS*: Korneev and O'Shea 2002). In principle, a duplication could also fortuitously combine sequences from two different functional sites to create a hybrid *cis*-regulatory region with a novel transcription profile. Gene duplications that persist are frequently followed by divergence in expression (Li and Noll 1994; Stauber et al. 2002; Gu et al. 2002) and may be followed by loss of complementary functional site modules (Ferris and Whitt 1979; Force et al. 1999). (5) *Gene conversion can spread regulatory elements within a gene family.* Examples from humans include growth hormone (Giordano et al. 1997), beta and gamma globins (Chiu et al. 1997; Patrinos et al. 1998), and MHC genes (Cereb and Yang 1994). Gene conversion is an ongoing process in RNA polymerase I-transcribed genes (which encode the 40S pre-rRNA that is processed to form 18, 5.8, and 28S rRNA) including associated transcriptional regulatory sequences, but not among the more heterogeneous RNA polymerase III-transcribed genes (White 2001). (6) *Sequences that have no prior function in regulating gene expression can become fortuitous functional sites.* In one case, gene duplications resulted in a former exon functioning in transcriptional regulation (*Sdic*: Nurminsky et al. 1998). A second example involves functional transcription factor binding sites within an exon (*nonA*: Sandrelli et al. 2001). These "hopeful monster" functional sites demonstrate that rare events can assemble functional *cis*regulatory sequences from seemingly unpromising material.

Changes in functional site sequence differ widely in their effects on transcription. Relatively little information exists about the functional consequences of naturally occurring differences in functional site sequences. A few studies have directly examined the biochemical impact of sequence differences on protein binding (e.g., Singh et al. 1998; Ruez et al. 1998; Wolff et al. 1999; Shaw et al. 2002). Most of what we know about functional consequences, however, comes from cases where the resulting transcription profile has been examined (e.g., Ross et al. 1994; Odgers et al. 1995; Tournamille et al. 1995; Belting et al. 1998; Indovina et al. 1998; Ludwig et al. 1998; Wang et al. 1999). In some of these cases, specific functional site sequence differences are correlated with phenotypic consequences; in most, however, the presence of multiple sequence differences makes it difficult to infer the precise basis for evolutionary changes in transcription. Divergence in functional site sequence and transcription profile are often poorly correlated: very similar functional sites can produce substantially different transcription profiles (e.g., parvoviruses: Storgaard et al. 1993; *TNFA* in primates: Haudek et al. 1998; *MMSP* in diptera: Christophides et al. 2000), while highly divergent functional site sequences can produce very similar transcription profiles (e.g., *runt* in *Drosophila*: Wolff et al. 1999; *brachyury* in ascidians: Takahashi et al. 1999; *yp* in *Drosophila*: Piano et al. 1999). The latter situation is not unusual. Indeed, many changes in functional site sequence do not alter transcription within the limits of experimental assays. Sequence changes might be functionally silent for several reasons. (1) *Substitutions and indels between transcription factor binding sites may not affect DNA-protein interactions.* It is probably generally true that nucleotides within binding sites are more functionally constrained than those that lie between binding sites. However, it is difficult to rule out the possibility that a supposed non-binding site nucleotide might in fact be part of an unrecognized binding site. Furthermore, small indels that do not directly involve binding sites may disrupt protein:protein interactions by placing proteins on opposite sides of the DNA helix or by changing

the spacing of binding sites. (2) *Changes in spacing between distant binding sites will be neutral in many cases.* Interactions among proteins associated with binding sites more than ~50 bp apart are probably mediated by DNA bending or looping, which may be largely insensitive to differences in spacing. (3) *Some within-consensus nucleotide substitutions in binding sites may be functionally neutral.* Certain changes in binding site sequence can preserve a particular DNA-protein interaction. Not all such changes will be neutral, however, as binding kinetics may differ, in turn altering transcription. Very little is known about the evolution of binding site consensuses, so sequence comparisons alone may be a poor guide as to which nucleotide substitutions within binding sites are likely to be functionally neutral. (4) *Eliminating an entire binding site may be functionally neutral.* Many functional sites contain multiple copies of the same binding site, raising the possibility of functional redundancy. Cases of probable binding site turn over (Ludwig and Kreitman 1995; Hancock et al. 1999; Piano et al. 1999; Liu et al. 2000; Dermitzakis and Clark 2002; Scemama et al. 2002) may have been possible because of functional redundancy. On the other hand, multiply represented binding sites within the same functional site are not always functionally redundant.

Gene expression profiles evolve frequently and in diverse ways. The literature of comparative gene expression has emphasized similarities and generally interpreted them as conserved features (DeRobertis and Sasai 1996; Holland and Holland 1999; Carroll et al. 2001). When comparing distantly related taxa, however, similarities in gene expression are often outweighed by apparently non-homologous features (Wray and Lowe 2000; Davidson 2001; Wilkins 2002). The abundance of population-level variation in functional site function means that expression differences could evolve quite rapidly under some conditions, and indeed substantial differences in gene expression can exist even between recently duplicated genes (Gu et al. 2002) or closely related species (Parks et al. 1988; Ross et al. 1994; Swalla and Jeffery 1996; Grbic et al.

1998; Kissinger and Raff 1998; Brunetti et al. 2001; Ferkowicz and Raff 2001). Several functional classes of evolutionary change in gene expression are evident. (1) *Changes in timing of gene expression*. Temporal changes have been documented from many taxa (e.g., Dickinson 1988; Wray and McClay 1989; Swalla and Jefferey 1996; Kim et al. 2000; Skaer et al. 2002). Heterochronies are a common pattern of anatomical evolution (McKinney and McNamara 1991), and must, at some level, involve heritable changes in the timing of gene expression. (2) *Changes in spatial extent of gene expression*. Many studies have found interspecific differences in the spatial extent of regulatory gene expression (e.g., Schiff et al. 1992; Abzhanov and Kaufman 2000; Brunetti et al. 2001; Scemama et al. 2002). Such changes are of particular interest when they affect regulatory genes, because of the relatively direct consequences for body proportions, organ size and number, and a great many other anatomical features. (3) *Changes in level of gene expression*. Evolutionary differences in transcription rate have also been documented (Regier and Vlahos 1988; Crawford et al. 1999; Wang et al. 1999). Such comparisons have been simplified with the advent of microarray technologies (e.g., Jin et al. 2001; Schadt et al. 2003). Because of this approach, we know more about differences in transcript abundance than any other kind of evolutionary change in gene expression. (4) *Changes in responsiveness of gene expression to external cues*. Evolutionary changes in transcriptional responses to physiological status, environmental conditions, and pheromones have also been documented (Brakefield et al. 1996; Cooper 1999; Abouheif and Wray 2002) . Such changes are a necessary component in the evolution of polyphenism and phenotypic plasticity and are therefore of considerable ecological interest. (5) *Sex-specific expression*. Evolutionary changes in differential gene expression among sexes have been documented (Schiff et al. 1992; Saccone et al. 1998; Christophides et al. 2000; Kopp et al. 2000) . Microarray surveys suggest that populations can harbor variation in which genes are expressed in a sex-specific manner (Jin et. al 2001).

(6) *Gains and losses of particular phases of gene expression* In multicellular organisms, many genes are expressed in a succession of spatially and temporally distinct phases during the life cycle (for examples see: Gerhart and Kirschner 1997; Carroll et al. 2001; Davidson 2001). A gene whose expression requires a particular transcription factor during a specific phase of expression may be "abandoned" by that regulator if it is no longer expressed in the appropriate region. Examples include several independent losses of patterning roles for homeodomain transcription factors in arthropods (Dawes et al. 1994; Falciani et al. 1996; Grbic et al. 1988; Mouchel-Vielh et al. 2002). Conversely, a new regulatory linkage may be established if a functional site acquires a binding site for a different transcription factor, a process known as recruitment or cooption (Duboule and Wilkins 1998; Wilkins 2002). Many likely cases have been identified (Lowe and Wray 1997; Saccone et al. 1998; Keys et al. 1999; Brunezti et al. 2001; reviewed in Wilkins 2002). Evolutionary gains and losses of particular phases of gene expression may be facilitated by the modular organization of functional sites.

Changes in gene expression differ widely in their effects on organismal phenotype. Functional site function has both a biochemical phenotype, the gene expression profile, as well as an organismal phenotype, involving features such as anatomy, physiology, life history, and behavior. These biochemical and organismal effects are evolutionarily dissociable to some extent, since some changes in gene expression appear to have no consequence for organismal phenotype. Such changes in gene expression are analogous to conservative amino acid replacements in a protein, many of which are likewise thought to have no impact on organismal phenotype (Kimura 1983; Gillespie 1991). Several cases are known where the timing or spatial extent of gene expression differs among species without any obvious phenotypic consequence (e.g., *gld* in *Drosophila*: Schiff et al. 1992, Ross et al. 1994; *Cy* gene family in sea urchins: Fang and Brandhorst 1996, Kissinger et al. 1998; *msp130* in sea urchins: Wray and Bely 1994). Although it may be difficult to demonstrate beyond any

doubt that a particular difference in transcription is phenotypically silent, the opposite case is easier to establish. Differences in gene expression have been linked to diverse aspects of organismal phenotype, including: (1) *anatomy* (Burke et al. 1995; Averof and Patel 1997; Stern 1998; Wang et al. 1999; Lettice et al. 2002); (2) *physiology* (Abraham and Doane 1978; Matsuo and Yamazaki 1984; Dudareva et al. 1996; Sinha and Kellogg 1996; Stockhaus et al. 1997; Segal et al. 1999; Lerman et al. 2003); (3) *behavior* (Trefilov et al. 2000; Saito et al. 2002; Caspi et al. 2002; Enard et al. 2002b ; Fang et al. 2002; Hariri et al. 2002); (4) *disease susceptibility* (Tournamille et al. 1995; Shin et al. 2000; Bamshad et al. 2002; Meyer et al. 2002); (5) *polyphenism* (Brakefield et al. 1996; Abouheif and Wray 2002); and (6) *life history* (Allendorf 1982, 1983; Anisimov et al. 2001; Streelman and Kocher 2002).

PHARMACOGENOMICS

Most research and tools used in molecular biology have focused on the detection of RNA and proteins, and the DNA sequences that encode these RNAs and proteins. The emphasis on protein coding regions of DNA is a major limitation to understanding how the genome is implicated in human health and disease, because the vast majority of regulatory DNA is very important to disease, yet specific sequences and their locations generally are not known. Accordingly, any information concerning the specific sequences and where they may be found would provide immeasurable benefits to molecular biology and would advance human medicine greatly.

The use of pharmacogenomics in clinical trials, including the identification of gene variants that influence clinical responses to drug, is becoming increasingly important (see e.g., U.S. Patent Publication No. 2001/0034023 A1, which is incorporated herein by reference in its entirety). This growing area of medicine enables more individualized, science-based treatment decisions. Other aspects of pharmacogenomics include predicting drug response (efficacy) and limiting side effect profiles. The ability to better predict drug

response would allow individualized pharmacotherapy that could increase the chance of selecting an optimal drug for each patient and could offer savings in both time and cost of care, and substantially improve a patient's long-term prognosis (see U.S. Patent Publication No. 2003/0108938 A1).

Many drugs or other treatments are known to have highly variable safety and efficacy in different individuals. A consequence of such variability is that a given drug or other treatment may be effective in one individual, and ineffective or not well-tolerated in another individual. For example, the PDR shows that about 45% of patients receiving Cognex (tacrine hydrochloride) for Alzheimer's disease show no change or minimal worsening of their disease, as do about 68% of controls (including about 5% of controls who were much worse). About 58% of Alzheimer's patients receiving Cognex were minimally improved, compared to about 33% of controls, while about 2% of patients receiving Cognex were much improved compared to about 1% of controls. Thus a tiny fraction of patients had a significant benefit. Response to many cancer chemotherapy drugs is even worse. For example, 5-fluorouracil is standard therapy for advanced colorectal cancer, but only about 20-40% of patients have an objective response to the drug, and, of these, only 1-5% of patients have a complete response (complete tumor disappearance; the remaining patients have only partial tumor shrinkage). Conversely, up to 20-30% of patients receiving 5-FU suffer serious gastrointestinal or hematopoietic toxicity, depending on the regimen (see U.S. Patent Application Publication 2001/0034023 A1).

Thus, administration of such a drug to an individual in whom the drug would be ineffective would result in wasted cost and time during which the patient's condition may significantly worsen. Also, administration of a drug to an individual in whom the drug would not be tolerated could result in a direct worsening of the patient's condition and could even result in the patient's death (U.S. Patent Publication No. 2001/0034023 A1). Thus, there is a need for stratifying patients in a clinical trial in a manner that provides a better predictor of clinical outcome that

do the categories that are commonly used (*e.g.*, racial, ethnic, gender, and/or geographic origin).

SUMMARY OF THE INVENTION

It has been shown that for some drugs, over 90% of the measurable intersubject variation in selected pharmacokinetic parameters has been shown to be heritable (U.S. Patent Publication No. 2001/0034023 A1). Thus, it would greatly advantageous to identify genetic indicators associated with disease that serve as markers for patient responsiveness to a drug and thus aid in stratifying patients in clinical trials. This application concerns, inter alia, the identification of genomic regulatory elements associated with disease, and methods for identifying and exploiting sequence variations in these genomic regulatory elements that account for interpatient variation in the diseased state and for the interpatient variation in responsiveness to therapeutic agents.

Accordingly, the present invention provides methods for stratifying a patient in a subgroup of a clinical trial for the prevention or treatment of a disease selected from the diseases listed in Column 12 of Table 1, comprising determining the genotype of a functional site corresponding to said disease, and stratifying the patient in a subgroup of a clinical trial according to said genotype. Alternatively (or additionally), the stratification can be based on the disease associated with the functional site as can be determined from the name of the gene with which the functional site is associated.

In certain embodiments, the functional site whose genotype is the basis for said stratification is selected by identifying, for example by expression profiling, genes whose expression is altered in diseased cells, and identifying functional sites corresponding to these genes that differ in sequence from the sequence of the functional sites listed in Table 1, *i.e.*, functional sites whose genotype is a marker of a disease of interest.

The present invention yet further provides an isolated polynucleotide comprising a sequence selected from the group consisting of: (a) a functional site sequence provided in Table 1, (b) a complement of a functional site sequence provided in Table 1, (c) a sequence consisting of at least 10 contiguous residues of a functional site sequence provided in Table 1, (d) a sequence that hybridizes to a functional site sequence provided in Table 1, under moderately stringent conditions, and (e) a sequence having at least 75%, 80%, 85%, 90% or 95% identity to a functional site sequence in Table 1; wherein said sequence is not flanked at its 5' end or its 3' end by greater than 200 nucleotides of sequences that are contiguous to the sequence in the human genome. In various embodiments, the sequence is not flanked at its 5' end or its 3' end by greater than 100, 50, 25, 10, 5 or zero nucleotides of sequences that are contiguous to the sequence in the human genome. In certain embodiments the sequence of the polynucleotide of the invention consists of the functional site sequence listed in Table 1. In certain embodiments, the polynucleotide of the invention does not comprise vector sequences. In other embodiments, the polynucleotide of the invention is at least 30%, 40%, 50%, 60%, 70%, 80%, or at least 90% pure.

The present invention yet further provides vectors comprising the polynucleotides of the invention, in which the polynucleotides is optionally operably linked to (*i.e.*, regulates or modulates the expression of) an open reading frame, for example an open reading frame encoding a reporter. (*e.g.*, alkaline phosphatase, β-galactosidase, neomycin phosphotransferase, chloramphenicol acetyltransferase, dihydrofolate reductase, hygromycin phosphotransferase, beta-glucoronidase, green fluorescent protein, and luciferase).

The present invention further provides a host cell comprising a vector of the invention. The host cell can be a prokaryotic or a eukaryotic cell.

The present invention further provides an isolated polynucleotide comprising a plurality of the polynucleotides of the invention. Such a polynucleotide can contain two or functional sites corresponding to the same gene,

e.g., to regulate the expression of an open reading frame of interest such that the expression reflects part or all the expression pattern or the gene to which the functional sites correspond.

The present invention further provides primer pairs, each primer being at least 12 nucleotides, more preferably at least 15 nucleotides, in length, which primer pairs are capable of amplifying all or a portion of a functional site listed in Table 1. Such primer pairs can be capable of amplifying at least 50 base pairs, at least 75 base pairs, at least 100 base pairs, at least 150 base pairs or at least 200 base pairs of a functional site listed in Table 1.

The present invention yet further provides sets of two or more primer pairs for amplification all or a portion of two more functional sites listed Table 1. Each set can represent two or more functional sites listed for the disease in Table 1.

The present invention further provides a non-human animal comprising a polynucleotide of the invention, for example a rat, a mouse, or a non-human primate.

The present invention provides methods of profiling the state of a patient with respect to a disease listed in Column 12 of Table 1, comprising determining the genotype of one or more functional sites associated with a disease listed in Column 12 of Table. In certain embodiments, determining the genotype of one or more functional sites associated with the disease is achieved by profiling the genomic regulatory regions of a nucleic acid isolated from or amplified from the patient using a positionally addressable array (where the identity of and location of the nucleic acid at each position on the array is known), for example as described in U.S. Application No. 10/375,404, which is incorporated by reference herein in its entirety. In other embodiments, determining the genotype of one or more functional sites associated with the disease is achieved by real-time PCR, for example as described in International Application No. PCT/US02/16967, which is incorporated by reference herein in its entirety. In yet other embodiments,

determining the genotype of one or more functional sites associated with the disease is achieved by PCR amplification of the functional site from the patient's genomic DNA and sequencing the resulting PCR product. In various embodiments, at least 2, 3, 5, 10, 12, 15, 20, 25, 30, 40, or 50 functional sites associated with the disease are profiled. In other embodiments, at least 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60% or 75% of the functional sites listed in Table 1 as associated with the disease are profiled.

In certain embodiments, the present invention provides a positionally addressable polynucleotide array comprising a plurality of different polynucleotides, each different polynucleotide (a) differing in nucleotide sequence, (b) being affixed to a substrate at a different locus, (c) being in the range of 10-1000 nucleotides in length, and (d) being complementary and hybridizable to a functional site listed in Table 1 or its complement, and wherein the loci at which said different polynucleotides are situated are at least 5%, 10% or 15% of the total loci of the array. In one embodiment, each different polynucleotide is greater than 30 nucleotides and is designed so as not to contain a sequence of in the range of 15-30 nucleotides that occurs in the genome of the organism from which the functional sites are identified greater than 10 times. In one mode of the embodiment, the array is a tiling array in which at least a portion of the polynucleotides are complementary and hybridizable to the same functional site or its complement. Optionally, such polynucleotides comprise overlapping sequences.

The present invention further provides positionally addressable polynucleotide arrays to which nucleic acids are hybridized, in which the polynucleotides affixed to the array and/or the nucleic acids hybridized to the array are enriched in sequences that are hybridizable to functional sites or their complements. Such arrays can be solid phase arrays or semi-solid phase arrays.

In certain embodiments, the present invention provides a positionally addressable polynucleotide array to which nucleic acids are hybridized, said array

comprising a plurality of different polynucleotides, each different polynucleotide (a) differing in nucleotide sequence and (b) being affixed at a different locus to a substrate, said nucleic acids being enriched in functional sites listed in Table 1 or their complements or fragments of the functional sites or complements of at least 10 base pairs, said nucleic acids being hybridized to one or more discrete loci on the array.

In other embodiments, the present invention provides a positionally addressable polynucleotide array to which nucleic acids are hybridized, said array comprising a plurality of different polynucleotides, each different polynucleotide (a) differing in nucleotide sequence, (b) being affixed at a different locus to a substrate, (c) being in the range of 10-1000 nucleotides in length, and (d) being complementary and hybridizable to a functional site listed in Table 1 or its complement, and wherein the loci at which said different polynucleotides are situated are at least 5%, 10% or 15% of the total loci of the array.

In other embodiments, the present invention provides a positionally addressable polynucleotide array to which nucleic acids are hybridized, said array comprising a plurality of different polynucleotides, each different polynucleotide (a) differing in nucleotide sequence, (b) being affixed at a different locus to a substrate, (c) being in the range of 10-1000 nucleotides in length, and (d) being complementary and hybridizable to a functional site listed in Table 1 or its complement, wherein the loci at which said different polynucleotides are situated are at least 5%, 10% or 15% of the total loci of the array; and wherein said nucleic acids are enriched in ACEs or fragments thereof of at least 10 base pairs.

In other various embodiments of the foregoing methods and compositions, polynucleotides comprising or hybridizable to functional site sequences or their complement or fragments of the functional sites or complements of at least 15, 20, 30 or 40 nucleotides represent at least 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, or 90% of the polynucleotides on a positionally addressable polynucleotide array, and the polynucleotides comprising

or hybridizable to functional site sequences listed in Table 1 or their complements (or fragments thereof of at least 15, 20, 30 or 40 nucleotides) represent at least 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, or 90% of said polynucleotides comprising or hybridizable to functional site sequences or fragments. Further, in various embodiments, the plurality of polynucleotides on a positionally addressible array is at least 100, at least 200, at least 300, at least 400, at least 500, at least 600, at least 800, at least 1,000, at least 5,000, at least 10,000 or at least 20,000 different polynucleotides.

In a specific embodiment, a nucleic acid sample of a patient (*e.g.*, a genomic DNA sample), or a sample derived therefrom, is sequenced in order to determine whether one or more functional sites listed in Table 1 are present. Preferably, the genotype of the functional site is determined, in order to see whether the particular sequence listed in Table 1, or a variant thereof, is present. The sequencing can be done by any method known in the art, *e.g.*, sequencing by hybridization (*e.g.*, using the hybridization methods and arrays described above).

In a specific embodiment, the functional sites (*e.g.*, of a patient) being sequenced comprise preferably at least 3 different functional sites, more preferably at least 5 different functional sites, more preferably at least 10 different functional sites, more preferably at least 20 different functional sites, and yet more preferably at least 50 different functional sites. In a preferred embodiment, a profile of functional sites contains primarily or exclusively functional sites associated with the same gene, functional sites associated with the same disease (*e.g.*, as set forth in Table 1), and/or functional sites associated with a group of related diseases (*e.g.*, cancers or cardiovascular disorders).

BRIEF DESCRIPTION OF THE OF THE DRAWING(S)

Figure 1. The position of a clone from PS008 which was confirmed by a QRT-PCR to be a DNaseI hypersensitive site. Clone #123456 was mapped to

a genomic location on chromosome 3 approximately 5 kb upstream of the transcriptional start site of the *ECT2* gene. Primers were designed for an amplicon encompassing the clone and the hypersensitivity measured by quantitative real time PCR (as described in the Examples section).

Figure 2. Creation of Reference DNA. Nuclei are digested with DNaseI to preferentially introduced double-stranded breaks into DNaseI hypersensitive sites. These sites are repaired and A-tailed so they can be ligated to a common biotinylated adaptor. Following fraction of the genome by digestion with *Nla*III the biotinylated DNA is separated on paramagnetic strepavidin coated beads, and the bulk of the genome (*Nla*III-*Nla*III fragments) washed away. The isolated DNA, which is enriched in hypersensitive sites, is ligated to a second adaptor (to allow PCR amplification downstream) and recovered from the beads by *Not*I-digestion.

Figure 3. Creation of Subtractive DNA. DNA isolated from DNaseI-digested nuclei is either digested with *Nla*III alone (to create PS005-Subtractant DNA) or split into aliquots and digested with *Pst*I, *Sph*I, *Nsi*I or *Sac*I and then pooled (to create PS008-Subtractant DNA). Digestion by all these enzymes generates ends with four nucleotide 3' overhangs which are resistant to ExonucleaseIII-digestion. Exonuclease III will digest double-stranded breaks introduced by digestion with DNaseI to render the fragments single stranded. Single-stranded fragments are subsequently digested by Mung Beam Nuclease (a 5' to 3' exonuclease). Following digestion the remaining fragments are biotinylated by the dual action of Terminal Transferase in the presence of Biotin-ddUTP and chemical labelling with photobiotin. The resultant populations of Subtractant DNAs are heavily biotinylated and depleted in hypersensitive sites.

DETAILED DESCRIPTION OF THE INVENTION

The expression of genes relies upon the coordinated activities of numerous regulatory networks, all of which ultimately exert their influence through

functional sites within genomic DNA. This set of functional sites may be referred to as the "regulome." These functional sites represent the key regulatory regions of genomic DNA and, thus, govern gene expression and all related biological processes, including, *e.g.,* cell proliferation, differentiation, development, and apoptosis. Furthermore, since the vast majority of diseases are polygenic and due to quantitative variation in gene expression/regulation, the vast majority of functional genetic mutations that cause or modulate disease will be found within functional sites of the regulome.

The unique characteristics and phenotype of a cell are largely dictated by the specific pattern of gene expression associated with the cell. Indeed, it is widely understood that different cell types express different genes and that gene expression changes in response to different biological cues and external stimuli, such as, *e.g.*, growth factors, cytokines, and drugs. Accordingly, a cell may be characterized by its specific pattern of gene expression and activity, which, in turn, may be identified based upon the specific functional sites active or present in the cell. The identification of functional sites present in a specific cell, including, *e.g.*, a disease cell or a cell treated with a specific drug, therefore, provides a novel and powerful means of characterizing or identifying a cell and observing changes in cellular behavior associated with a variety of factors, including disease and drug treatment, for example.

The present invention provides novel compositions comprising functional sites identified in genomic DNA and methods for using the same. The functional sites of the present invention are listed in Table 1, along with the following identifying characteristics:

Table Legend: Sequences of functional site clones of the described HS-enriched libraries. The genomic location of each sequence is shown as is the closest gene and any association that gene has with a disease state.

Column 1: **SEQ ID NO:** of the functional site in the sequence listing.

Column 2: **CID**. Candidate Identity; a unique number defining each sequence.

Column 3: **Library**. The HS-enriched library out of which the sequence was derived.

Column 4: **Chromosome**. The number of the human chromsome to which the sequence was mapped using Build 12 of the human genome.

Column 5: **Position**. The genomic co-ordinates of the sequence (reading from the 5' end) following mapping to the assigned chromosome on UCSC build 12 of the human genome (UCSC version hg12, obtained from the UCSC Genome Bioinformatics Site, located at genome.ucsc.edu).

Column 6: **Gene Name**. The closest gene to the position of the functional site.

Column 7: **Feature**. The sequence is mapped relative either upstream of the 5' end of the chosen gene (**5'**), is internal to the gene and mapped downstream to the 5' end but not beyond the 3' end of the gene (**TRANSC**) or downstream of the 3' end of the gene (**3'**).

Column 8: **Offset**. The position of the sequence relative to the indicated Feature.

Column 9: **Sense**. The strand on which the sequence has been read; upper strand (1) or lower strand (-1).

Column 10: **Gene ID**. The identity of the selected gene as provided by ENSEMBL.

Column 11: **Gene Description**. The information regarding the selected gene as appears in the associated GeneCards or Ensembl Database.

Column 12: **Disease**. Disease states were as listed in the GeneCards Database (http://bioinfo.weizmann.ac.il/cards/).

Column 13: **Sequence**. The sequence pertaining to the CID.

Such compositions and methods allow the identification and characterization of functional sites present within different cells and tissues,

including disease cells, and the identification and characterization of cells and cellular responses. The compositions and methods of the invention provide an integrated approach combining molecular, high throughput, and bioinformatic and computational methods, which permits genome-wide global analysis of functional sites. Such genome-wide profiling of functional sites has broad applications in cell characterization, and may be applied, *e.g.,* to identify disease genes and regulatory networks, determine the effects of drugs and other agents, and develop unique characteristic markers of cells, including different cell or tissue types, disease cells, and cells treated with different drugs or agents, for example.

The invention, in certain embodiments, provides functional sites and libraries and arrays of functional sites. Relatedly, the invention provides, *inter alia,* methods of identifying or profiling functional sites within cells, methods of identifying and characterizing cells, and methods of regulating gene expression, as further described *infra*.

The following definitions are provided to assist in understanding the various embodiments of the invention as described:

A "functional site" or an "active chromatin element" or "ACE" (which terms are used interchangeably herein) is a specific region of genomic DNA, which in the context of nuclear chromatin, is associated with a disruption in chromatin structure and is accessible to a DNA-modifying agent, and which is associated with preferably one of the following characteristics: (i) 10-50 times greater hypersensitivity to the DNA modifying agent relative to the nearby region; (ii) 50-100 times greater hypersensitivity to the DNA modifying agent relative to the nearby region; (iii) 100-150 times greater hypersensitivity to the DNA modifying agent relative to the nearby region; and (iv) 150-200 times greater hypersensitivity to the DNA modifying agent relative to the nearby region; and/or one, two, three, four, five, six or all seven of the following characteristics: (v) an intrinsic ability to confer hypersensitivity to the DNA modifying agent when excised from its native location and inserted into at least one different location in the genome of a cell of

the same cell type ; (vi) the ability to reconstitute a site that is hypersensitive to the DNA modifying agent when a nucleic acid comprising the nucleotide sequence flanked by at least 1000 bp on each side is assembled into chromatin in an in vitro reconstitution assay in the presence of nucleosomal proteins and a cell extract; (vii) is non-nucleosomal when present in chromatin isolated from one or more cells; (viii) is embedded in DNA associated with histones that have a high degree of acetylation when present in chromatin isolated from one or more cells; (ix) greater solubility than nucleosomal material in moderate salt solutions (e.g., 150 mM NaCl and 3mM $MgCl_2$) when present in chromatin isolated from one or more cells; (x) is a non-coding sequence; or (xi) does not occur greater than 10 times in a genome of the organism in which the ACE is identified.. Functional sites include isolated polynucleotides corresponding to and forming an inseparable and dominant component of functional sites. Functional sites are biologically-bounded by flanking nucleosomes and span the inter-nucleosomal interval, which is approximately 150-250 base pairs in length. A functional site typically contains a core domain of approximately 80-100 base pairs in length, which is required for formation of the functional site *in vivo*. In addition, a functional site sequence may further contain flanking regions that modulate the activity of the core domain. A functional site may also be referred to herein as an active chromatin element or ACE.

A "functional site variant" is a region of genomic DNA, which differs in sequence as compared to a functional site at the same genomic location. A functional site variant may or may not be a functional site in one or more cells wherein the corresponding functional site is present.

A " chromatin modifying agent" (CMA) is an agent capable of modifying genomic DNA, in the context of nuclear chromatin, in a detectable manner. Examples of DNA-modifying agents and associated modifications include nucleases (non-specific, *e.g.*, DNase I, and sequence-specific, *e.g.,* restriction endonucleases), DNA-binding proteins (modified and non-modified), DNA-

modifying enzymes (*e.g.,* methyl transferases, acetylases), DNA-intercalating agents (*e.g.,* bleomycin, topoisomerases), and integrating viruses.

The "regulome" is the complete set of all functional sites present in a species.

A "tissue regulome" is the complete set of all functional sites present in a particular cell or tissue.

A "regulotype" is a set of functional sites present in a particular individual or organism. Thus, a "regulotype" is specific for the particular individual or organism.

A "tissue regulotype" is a set of functional sites present in a particular cell or tissue of a particular individual or organism. Thus, a tissue regulotype is specific for the particular cell or tissue-type.

"Profiling" is identifying the presence or absence of functional sites in a particular cell at one or more particular genomic loci. Depending upon the origin and/or treatment of the cell being profiled, profiling includes, *e.g.,* tissue profiling, disease profiling, drug profiling, and functional mutant profiling. Profiling may be used to determine the pattern of functional site presence or absence specific to a particular cell or tissue, including, *e.g.,* a diseased cell or a cell treated with a drug.

"Locus profiling" is identifying functional sites present in a particular cell at a particular genomic locus.

A "gene" is a contiguous region of genomic DNA that consists of the sequences that encode a polypeptide and substantially all of the sequences that regulate expression of the coding sequences.

A "regulatory pathway" is a collection of cellular constituents that regulate the expression of one or more gene products, wherein each cellular constituent is influenced according to some biological mechanism (*e.g.,* cooperative binding, DNA or protein modification, etc.) by one or more other constituents of the collection.

An "array" is a plurality of different nucleic acids immobilized at positionally-addressable locations on a solid phase surface.

A "microarray" is an array in which the immobilized nucleic acids are located within a region of less than 6.25 cm^2 in size (although the solid phase surface can be much larger).

A "regulatory array" is an array of nucleic acids, each comprising a functional site sequence or functional site variant sequence.

A "pharmaceutical regulatory array" is an array of nucleic acids, each comprising a functional site sequence or functional site variant sequence associated with one or more specific genes known or presumed to be involved in pharmaceutical response or metabolism.

A. Polynucleotides, Vectors, and Cells Comprising Functional Sites

Until now, sequences and location information of functional sites within genomic DNA have not been available on a large scale and medical advances have been greatly hindered from this lack of information. The sequences of the discovered fragments, herein termed "functional sites" and their genomic locations, are provided in Table 1. These sequences are typically about 100 to 300 bases long, and it is estimated that they can each bind approximately 6 - 10 proteins. Genomic locations include the chromosomal location of a nucleic acid sequence, as identified by routine chromosomal mapping procedures or by comparison to a database of nucleic acid sequences and their chromosomal location, for example. It is estimated that there are approximately 150,000 - 200,000 functional sites present within the human genome and that approximately 40,000 - 50,000 functional sites are active in any particular cell type.

The DNA sequences listed in Table 1 are provided with their genome locations. These functional site sequences were discovered as hypersensitive sites in chromatin. While generally not including repetitive sequences or encoding protein sequences, each functional site sequence listed, nevertheless, was found

morphologically active in certain human cells. The functional activity of each functional site is important to the expression of one or more protein encoding genes.

1. Functional Site Polynucleotides

In one embodiment, the invention provides polynucleotides comprising, consisting essentially, or consisting of one or more functional sites, or variants or complements thereof. Specific functional site polynucleotide sequences of the invention are provided in Table 1. The polynucleotides of the invention are not, however, limited to these specific and illustrative sequences. Rather, the invention encompasses any and all functional sites of any and all genomes. For example, functional sites of the present invention include those identified or present in the genome of any animal, virus, or plant. In certain embodiments, functional sites include those present in a mammalian genome, such as, for example, a human, mouse, or pig genome.

i. Size

Functional site sequences are generally size-restricted and biologically bounded by (1) the positions of flanking nucleosomes and (2) limits on the area of DNA over which thermodynamically stable nucleoprotein complexes may form. The extent of the functional site typically spans the inter-nucleosomal interval of approximately 150-250 bp. This interval corresponds to the size of sequence that is needed to place a nucleosome, and it has been a common assumption that functional sites represent a break in the cannonical nucleosomal array that constitutes the vast majority of chromatin.

In certain embodiments, a core domain within a functional site sequence can be identified which is restricted to a region of approximately 80-100 base pairs in length, over which DNA-protein interactions take place. It has been shown that the cooperative binding of transcription factors to such core regions are

sufficient to exclude a nucleosome *in vitro* (Adams and Workman, Mol. Cell Biol., 15: 1405), and this has been accepted as a common mechanism for how these sites may form *in vivo*. Nucleosomal mapping experiments have shown that functional sites such as the *Drosophila hsp26* promoter (Lu *et al.*, EMBO J. 14; 4738) and the human β-globin HS2 (Kim and Murray, Int. J. Biochem. Cell Biol 33: 1183) are non-nucleosomal. It is thought that most functional sites are non-nucleosomal in nature (Boyes and Felsenfeld, EMBO J. 15: 2496; Wallrath *et al.*, Bioessays 16:165). These conclusions are well-supported in the literature (e.g., *ibid* and Struhl K. Science. 2001 Aug 10;293(5532):1054-5). However, several functional sites are known to still have bound histone proteins and transcription factors, suggesting that the functional sites may exist in conjunction with a modified nucleosome.

Flanking sequences surrounding the core region appear to modulate the activity of this core region, though this effect tapers off as the distance from the core region increases. The boundaries of the sequences needed for functional activity, *e.g.,* hypersensitivity activity, can be defined functionally by performing deletional analysis in studies following stable transfection of cells (Philipsen *et al.*, EMBO J. 9: 2159) or transgenic studies (Zhou et al., J Cell Sci. 108:3677). These approaches define the minimum extent of sequence required to retain the biological function associated with the functional site under examination.

ii. Clusters of transcription factors binding elements

High resolution studies of DNA sequences of known regulatory regions demonstrate that these regions often represent clusters of recognition sites for promoter-specific DNA-binding proteins (Emerson *et al.*, 1985). Very few of these binding elements can be predicted on the basis of DNA sequence alone. Recent studies using chromatin immunoprecipitation have revealed that the 'consensus' binding motifs of transcription factors have both low sensitivity and very low specificity in predicting actual sites of *in vivo* DNA-protein interaction.

However, this prediction can be substantially improved (and in many cases rendered definitive) with prior knowledge that the motif occurs in a region known to comprise a functional site.

iii. Catalytic activity

Functional site-forming genomic DNA sequences have unique physical properties. In principle, these sequences can be said to function in a 'catalytic' manner that is analogous to the interaction between an enzyme and its substrate. These DNA sequences contribute to the free energy of formation of a nucleoprotein complex in a manner that dramatically increases its probability of activation vs. neighboring DNA regions.

An important finding has been that these sequences only function so when they are assembled into genomic chromatin. The sequences adopt a particular topological confirmation, which is compatible with the coalescence of numerous proteins, some in contact with DNA and some in contact with other proteins. This results in the formation of a nucleoprotein complex. The formation of the complex is precisely correlated with a particular sequence, which drastically lowers its activation energy with respect to other sequences, and also with respect to contact of those proteins with one another *in vivo* under random circumstances. The final product is stochastic, in the sense that it forms in an all-or-none fashion (*e.g.,* Felsenfeld *et al* Proc Natl Acad Sci U S A. 1996 Sep 3;93(18):9384; Boyes & Felsnfeld EMBO J. 1996 May 15;15(10):2496).

The rate of formation can be measured through interrogation with the quantitative nucleosensitivity assay described below and in more detail in PCT Publication No. WO 02/097135 and U.S. Patent Applications Serial No. 10/157,027 and Serial No. 10/319,440, which are hereby incorporated be reference in their entirety. When examined over a time-course of digestion, a characteristic 'signature' relationship can be derived for each catalytic sequence, which can be quantified and assigned a mathematical constant. A further conceptual parallel

with other catalytic processes is that nucleoprotein complex formation can be manipulated through the introduction of point mutations or small deletions or insertions in the "active site" (critical DNA binding bases) or "allosteric" sites (juxtaposed sequences). This principle has been demonstrated in numerous publications (*e.g.,* Stamatoyannopoulos *et al* EMBO J. 1995 Jan 3;14(1):106).

iv. Intrinsic ability to form

A further defining feature of functional sites is that the function of the DNA sequence component – *i.e.* its complex-forming activity – is intrinsic. The principal evidence for this is the fact that these sequences can be excised and inserted into other positions in the genome, where they exhibit the same functional chromatin activities. Substantial experimental experience from model systems has revealed that functional sites can form when included in either constructs used to create stably transfected cell lines (Fraser *et al.*, 1990) or transgenic animals (Lowrey *et al.* Proc Natl Acad Sci U S A. 1992 Feb 1;89(3):1143-7; Levy-Wilson *et al.*, 2000).

v. Activity in transgenic systems

Many functional sites can be shown to have regulatory influences on the expression of reporter genes when included in constructs in transfection or transgenic systems. Such systems can be used to demonstrate activities associated with promoters (Furbass *et al.*, 2001), transcriptional enhancers (Levy-Wilson *et al.*, 2000) and transcriptional silencers (Oritz *et al.*, 1999). Functional sites have also been reported to behave as insulator elements, defined as sequences that prevent the transmission of chromatin structure features associated with the genomic location into which the construct has integrated, in various transgenic models (Li *et al.*, 2002; Mustkov *et al.,* 2002; Rivella *et al.*, 2000). Functional sites can act as elements capable of opening chromatin, which may act singly (Nemeth *et al.*, 2001) or in a coordinated fashion with other

functional sites (commonly termed a Locus Control Region (Li *et al.,* 2002; Shewchuk *et al.*, 2001)).

As such, these transgenic assays represent a tool for identifying and classifying functional sites on the basis of function and also defining the minimum size of fragment on which the function is confined.

vi. Activity in chromatin reconstitution systems

Functional sites can be included in templates for reconstitution protocols (Leach *et al.,* 2002) or *in vitro* assembly systems (Becker *et al.*, 1991) and are capable of directing the formation of chromatin structure similar to that detected *in vivo.*

vii. Nucleoprotein complexes

In general, the majority of functional sites are believed to bind multiple (*e.g.,* three or more – with an expected average of 6-7) DNA binding proteins, which may be, *e.g.,* either ubiquitous transcription factors or proteins with a specific pattern of expression. The cooperative binding of transcription factors has been shown to be sufficient to exclude a nucleosome *in vitro* (Adams and Workman, 1995), and this has been accepted as a common mechanism for how these sites may form *in vivo.* Nucleosomal mapping experiments have shown that functional sites such as the *Drosophila hsp26* promoter (Lu *et al.*, 1995) and the human β-globin HS2 (Kim and Murray, 2001) are non-nucleosomal. It is thought that most functional sites are non-nucleosomal in nature (Boyes and Felsenfeld, 1996; Wallrath *et al.*, 1994).

It has also been proposed and demonstrated that, in certain rare circumstances, some DNA sequences can form functional sites in the absence of protein binding (*i.e.,* purely on the basis of their internal structural properties). Examples of these include the CpG-island associated with the human glucose-6-phosphate dehydrogenase gene that forms in yeast (Mucha *et al.*, 2000) and

sequences associated with repeats giving rise to human chromatin fragile sites (Hsu and Wang, 2002). Other functional sites have been identified in ternary complexes between the bound transcription factors, underlying DNA sequence and the still associated histones (Steger and Workman, 1997).

viii. Fractionation properties

Typically, functional sites are embedded in accessible chromatin. Some of the discovered properties of accessible transcriptionally competent chromatin include increased generalized sensitivity to nuclease digestion, patterns of histone modification (accessible chromatin has high levels of histone acetylation) and higher solubility in moderate salt solutions (such as 150 mM NaCl and 3 mM $MgCl_2$). These properties allow the preparation of chromatin fractions enriched in functional sites (Spencer and Davie, 2001).

ix. Biological activities

Focal alterations in chromatin structure, such as those associated with functional sites, are the hallmark of active regulatory sequences in eukaryotic genomes. These alterations display remarkably similar physical properties irrespective of genomic location or even of species of origin. Exemplary activities are provided in Table 2

Table 2. Activities Associated with Functional Sites

Property	*Definition*	*Examples*	*Reference*
Promoter	Transcriptional promoter	Murine retroviral MMTV-LTR	Bresnick *et al.*, 1992
		Drosophila hsp26	Keene *et al.*, 1981
		Human cholinergic gene	Tanaka *et al.*, 1998
Transcriptional Enhancer	Upregulates transcription from linked gene	Human β-globin HS2	Kong *et al.*, 1997
		Human apoB enhancer	Levy-Wilson *et al.*, 2000
		Rat PPK enhancer	Holloway and La Gamma,
		Human CD34 enhancer	

			1992 Radomska *et al.*, 1998
Transcriptional Silencer	Downregulates transcription from linked gene	Mouse Igκ silencer Chicken apoVLDLII silencer Murine IL-5 silencer Murine TCRα silencer	Liu *et al.*, 2002 Hache and Deeley, 1988 Arima *et al.*, 2002 Oritz *et al.*, 1999
Matrix Attachment Region	Tether chromatin to protein backbone	MARs within human CD8 gene complex Chicken α-globin MAR	Kieffer *et al.*, 2002 de Moura Gallo *et al.*, 1992
Origin of Replication (ORI)	Origin of DNA replication	*Puff II/9A* ORI	Urnov *et al.*, 2002
Recombination Sites	Sites of frequent chromosome translocations	AML1/RUNX1 breakpoints in t(8;21) leukemia	Zhang *et al.*, 2002

x. Position relative to genes

An important feature of functional sites that has emerged (and, in some cases such as the globin genes, has been exhaustively investigated) is that the genomic proximity of a gene to a functional site is the principal determinant of the influence of that functional site on the regulation of that gene. Functional site sequences may be located upstream (5'), downstream (3') or within genomic regions containing transcribed regions of a gene. Accordingly, functional sites may be located within transcribed regions of a gene.

Functional sites may also be located in clusters within a region of genomic DNA. Each individual functional site is typically involved in the regulation of one or more genes. However, clusters or combinations of functional sites often coordinately regulate genes. That is, it was found that many functional sites can work together, as will be appreciated by a skilled artisan. Many of these combinations are seen as clusters physically located on the same chromosome or near a certain gene, for example. However, other functional sites coordinately control expression, even though they are found in disparate regions of the genome. These groups may be identified by assays that detect their effects, such

as assays that compare whether the functional sites of the invention are active in particular cell types or under particular conditions such as growth conditions or chemical or environmental exposures. Functional sites that are present or active in the same or similar cells or conditions are likely involved in the coordinate regulation of one or more genes. Accordingly, in certain embodiments, the invention provides sets of functional sites associated with a particular gene or cluster. Such functional sites may be associated with a specific chromosome, and may be within a specific distance from each other, including, for example, within 50 bp, 100 bp, 500 bp, 1 kb, 2 kb, 5 kb, 10 kb, 100 kb, or greater than 100 kb.

xi. **Repetitive content**

Functional site sequences can essentially be thought of as being unique in the genome, save in cases where the sequences lie in segmental duplications.

xii. **Method of identifying**

Functional sites may also be defined or characterized based upon their method of identification. Detailed methods of identification are described below, and in certain embodiments, functional sites of the invention include those sequences identified according to any one of these methods. In certain embodiments, functional sites are genomic sequences that are accessible to or modified by any DNA modifying agent, including those described *infra*.

b. **Subsets and Combinations of Functional Sites**

In certain embodiments, the invention includes sets or groups of functional sites. These sets may be characterized by any means available, including, for example, the specific DNA cleaving or tagging agent used to identify the functional sites, the specific cell or tissue source of genomic DNA from which

the functional sites were isolated, or the genomic location of the functional sites, for example.

In certain embodiments, the invention identifies and includes functional sites identified from a specific tissue or cell. Further, these functional sites may be limited to those identified at a specific or identifiable biological point or condition, such as, for example a certain developmental stage, cell cycle state or diseased state. Accordingly, the present invention includes sequences comprising functional sites, or fragments or portions thereof, identified in the genome of specific cells or tissues. By identifying functional sites present in a particular cell type and/or at a specific biological condition, the invention provides a discrete genomic fingerprint, referred to as a "tissue regulotype" associated with the specific cell or tissue, which may be used to identify cells and identify genes that govern a variety of cellular processes, including, for example, cellular differentiation, specialized cell function, and/or disease establishment and/or progression.

In certain embodiments, the invention includes only newly identified functional sites or sequences.

In another embodiment, the invention includes the polynucleotide sequences of genes identified as being regulated by functional sites of the invention, their corresponding cDNAs and complements, primers specific for the genes or cDNAs; polypeptides encoded by the genes, and antibodies specific for these encoded polypeptides.

The invention further includes combinations and groupings of functional sites. Each individual functional site is involved in the regulation of one or more genes. However, combinations of functional sites typically coordinately regulate genes, as will be appreciated by a skilled artisan. Many of these combinations are seen as clusters physically located on the same chromosome or near a certain gene, for example. However, other functional sites coordinately control expression, even though they are found in disparate regions of the

genome. These groups may be identified by assays that detect their effects, such as assays that compare whether the functional sites of the invention are active in particular cell types or under particular conditions such as growth conditions or chemical or environmental exposures. Functional sites that are present or active in the same or similar cells or conditions are likely involved in the coordinate regulation of one or more genes. Accordingly, in certain embodiments, the invention provides arrays of functional sites associated with a particular gene or cluster. Such functional sites may be associated with a specific chromosome, and may be within a specific distance from each other, including, for example, within 100 bp, 500 bp, 1 kb, 2 kb, 5 kb, 10 kb, 100 kb, or greater than 100 kb.

In one embodiment, the invention provides a fusion polynucleotide consisting of or comprising a plurality of functional site sequences. In certain embodiments, a polynucleotide consisting of or comprising a plurality of functional site sequences includes multiple functional site sequence isolated according to a procedure described herein and concatamerized to form a single polynucleotide. In one embodiment, the polynucleotide may contain sequences corresponding to all or part of a restriction enzyme recognition site or linker sequence between each previously isolated functional site sequence. Fusion polynucleotides according to the invention may contain specific sets of functional sites, such as those associated with a specific cell type, disease, or drug treatment, for example. Fusion polynucleotides of the invention represent portions of the genome corresponding to functional sites. Fusion polynucleotides range in length and may, in certain embodiment, contain greater than 10 megabases (mb). Accordingly, the invention includes fusion polynucleotides of at least 1 kb, at least 5 kb, at least 10 kb, at least 50 kb, at least 100 kb, at least 500 kb, at least 1 mb, at least 2 mb, at least 3 mb, at least 4 mb, at least 5 mb, at least 6 mg, at least 7 mg, at least 8 mb, at least 9 mb, at least 10 mb, and all integer values in between. The invention further comprises functional fragments of fusion polynucleotides, which contain

one or more functional sites or core regions. Such fragments are described more generally *infra*.

c. Complements, Variants and Fragments of Functional Sites

The invention also includes polynucleotides comprising variants and complements of polynucleotide sequences of the invention. Complements may be used for a variety of purposes, including, for example, to detect the presence of a functional site sequence. In certain embodiments, complements are completely complementary to a polynucleotide sequence of the invention, including fragments thereof. However, the skilled artisan would understand that it is not required that complements are completely complementary to the entirety of a polynucleotide of the invention. In certain embodiments, complements are complementary to a portion of any polynucleotide of the invention and may be less than completely complementary. In specific embodiments, however, complements of the invention are capable of hybridizing to a polynucleotide of the invention under stringent or moderately-stringent conditions, as set forth below. As such, complements include oligonucleotides, such as those suitable for performing polymerase chain reaction.

The invention includes variants of polynucleotides of the invention and complements thereof. Examples of specific variants include allelic variants, including those associated with a disease and homologs from different organisms or species. Typically, polynucleotide variants will contain one or more substitutions, additions, deletions and/or insertions. Variants also encompass homologous genes of xenogenic origin.

The invention includes variants lacking one or more functions associated with the corresponding functional site of the invention, *e.g.* the ability to bind a polypeptide bound by the functional site, the ability to regulate gene expression in the same manner as the functional site, or the ability to be identified

according to the procedures described herein to identify functional sites. In certain embodiments, a variant is associated with a disease.

In other embodiments, variants retain one or more functions associated with the corresponding functional site. Functional sites of the invention typically form nucleoprotein complexes by binding one or more proteins. The skilled artisan would recognize that such binding may not require the exact sequence of a functional site of the invention and that certain nucleotide deletions, additions, or substitutions may be tolerated without substantially or completely preventing binding. Indeed, it has been shown that protein binding nucleic acid sequences frequently comprise a consensus sequence, which may consist of the core nucleotides required for protein binding. Accordingly, functional variants of the invention include polynucleotides with an altered sequence as compared to an identified functional site, but which retain one or more physical or functional properties of the functional site, including any of the propertied described above, the ability to affect transcription of a linked gene, or the ability to bind the same polypeptide as the native sequence, for example. Such binding may be determined by any method available in the art, including, for example, electrophoretic mobility shift assays performed in the presence or absence of an antibody specific for the polypeptide that binds the native polynucleotide.

Variants of the invention may be identified by a variety of means, including sequence homology to a polynucleotide of the invention or the ability to hybridize to a polynucleotide sequence of the invention or complement thereof. In certain embodiment, the invention includes polynucleotides with at least 60% identity, at least 70% identity, at least 80% identity, at least 90% identity, at least 95%, at least 98%, at least 99%, or any integer value between and including 70% and 99% identity, to a polynucleotide of the invention, including a functional site or fragment or complement thereof.

The term sequence homology, as described herein, refers to the sequence relationships between two or more nucleic acids, polynucleotides,

proteins, or polypeptides, and is understood in the context of and in conjunction with the terms including: (i) reference sequence, (ii) comparison window, (iii) sequence identity, (iv) percentage of sequence identity, and (v) substantial identity or homologous.

(i) A reference sequence refers to a sequence used as a basis for sequence comparison. A reference sequence may refer to a subset of or the entirety of a specified sequence or complement thereof.

(ii) A comparison window includes reference to a contiguous and specified segment of a polynucleotide sequence, wherein the polynucleotide sequence may be compared to a reference sequence and wherein the portion of the polynucleotide sequence in the comparison window may comprise additions, substitutions, or deletions (*i.e.*, gaps) compared to the reference sequence (which does not comprise additions, substitutions, or deletions) for optimal alignment of the two sequences. Generally, the comparison window is at least 20 contiguous nucleotides in length, and optionally can be 30, 40, 50, 100, or longer. Those of skill in the art understand that to avoid a misleadingly high similarity to a reference sequence due to inclusion of gaps in the polynucleotide sequence a gap penalty is typically introduced and is subtracted from the number of matches.

Methods of alignment of sequences for comparison are well known in the art. Optimal alignment of sequences for comparison may be conducted by the local homology algorithm of Smith and Waterman, *Adv. Appl. Math.* 2: 482 (1981); by the homology alignment algorithm of Needleman and Wunsch, *J. Mol Biol.* 48: 443 (1970); by the search for similarity method of Pearson and Lipman, *Proc. Natl. Acad. Sci.* 8: 2444 (1988); by computerized implementations of these algorithms, including, but not limited to: CLUSTAL in the PC/Gene program by Intelligenetics, Mountain View, California, GAP, BESTFIT, BLAST, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group (GCG), 7 Science Dr., Madison, Wisconsin, USA; the CLUSTAL program is well described by Higgins and Sharp, *Gene,* 73: 237-244, 1988; Higgins and Sharp, CABIOS

:11-13, 1989; Corpet, *et al., Nucleic Acids Research,* 16:881-90, 1988; Huang, *et al., Computer Applications in the Biosciences* 8:1-7, 1992; and Pearson, *et al., Methods in Molecular Biology* 24:7-331, 1994. The BLAST family of programs which can be used for database similarity searches includes: BLASTN for nucleotide query sequences against nucleotide database sequences; BLASTX for nucleotide query sequences against protein database sequences; BLASTP for protein query sequences against protein database sequences; TBLASTN for protein query sequences against nucleotide database sequences; and TBLASTX for nucleotide query sequences against nucleotide database sequences. See, *Current Protocols in Molecular Biology,* Chapter 19, Ausubel, *et al.,* Eds., Greene Publishing and Wiley-Interscience, New York, 1995. New versions of the above programs or new programs altogether will undoubtedly become available in the future, and can be used with the present invention.

Unless otherwise stated, sequence identity/similarity values provided herein refer to the value obtained using the BLAST 2.0 suite of programs using default parameters. Altschul *et al., Nucleic Acids Res,* 2:3389-3402, 1997. It is to be understood that default settings of these parameters can be readily changed as needed in the future.

(iii) "Sequence identity" or "identity" in the context of two nucleic acid or polypeptide sequences includes reference to the residues in the two sequences which are the same when aligned for maximum correspondence over a specified comparison window, and can take into consideration additions, deletions and substitutions.

(iv) "Percentage of sequence identity" means the value determined by comparing two optimally aligned sequences over a comparison window, wherein the portion of the polynucleotide sequence in the comparison window may comprise additions, substitutions, or deletions *(i.e.,* gaps) as compared to the reference sequence (which does not comprise additions, substitutions, or deletions) for optimal alignment of the two sequences. The

percentage is calculated by determining the number of positions at which the identical nucleic acid base or amino acid residue occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the result by 100 to yield the percentage of sequence identity.

(v) (i) The term "substantial identity" or "homologous" in their various grammatical forms means that a polynucleotide comprises a sequence that has a desired identity, for example, at least 60% identity, preferably at least 70% sequence identity, more preferably at least 80%, still more preferably at least 90% and most preferably at least 95%, compared to a reference sequence using one of the alignment programs described using standard parameters. One of skill will recognize that these values can be appropriately adjusted to determine corresponding identity of proteins encoded by two nucleotide sequences by taking into account codon degeneracy, amino acid similarity, reading frame positioning and the like. Substantial identity of amino acid sequences for these purposes normally means sequence identity of at least 60%, more preferably at least 70%, 80%, 90%, and most preferably at least 95%. It further includes sequences with at least 70-99% sequence identify, including all integer values in-between, including, for example, 90, 91, 92, 93, 94, 95, 96, 97, and 98.

Another indication that nucleotide sequences are substantially identical is if two molecules hybridize to each other under stringent conditions. The phrase "stringent hybridization conditions" refers to conditions under which a probe will hybridize to its target complementary sequence, typically in a complex mixture of nucleic acids, but to no other sequences. Stringent conditions are sequence-dependent and circumstance-dependent; for example, longer sequences hybridize specifically at higher temperatures. An extensive guide to the hybridization of nucleic acids is found in Tijssen, *Techniques in Biochemistry and Molecular Biology-Hybridization with Nucleic Probes,* "Overview of principles of hybridization and the strategy of nucleic acid assays" (1993). In the context of the

present invention, as used herein, the term "hybridizes under stringent conditions" is intended to describe conditions for hybridization and washing under which nucleotide sequences at least 60% homologous to each other typically remain hybridized to each other. Preferably, the conditions are such that sequences at least about 65%, more preferably at least about 70%, and even more preferably at least about 75% or more homologous to each other typically remain hybridized to each other.

Generally, stringent conditions are selected to be about 5-10°C lower than the thermal melting point (Tm) for the specific sequence at a defined ionic strength pH. The Tm is the temperature (under defined ionic strength, pH, and nucleic concentration) at which 50% of the probes complementary to the target hybridize to the target sequence at equilibrium (as the target sequences are present in excess, at Tm, 50% of the probes are occupied at equilibrium). Stringent conditions will be those in which the salt concentration is less than about 1.0 M sodium ion, typically about 0.01 to 1.0 M sodium ion concentration (or other salts) at pH 7.0 to 8.3 and the temperature is at least about 30°C for short probes (for example, 10 to 50 nucleotides) and at least about 60°C for long probes (for example, greater than 50 nucleotides). Stringent conditions may also be achieved with the addition of destabilizing agents, for example, formamide. For selective or specific hybridization, a positive signal is at least two times background, preferably 10 times background hybridization.

Exemplary, non-limiting stringent hybridization conditions are as following: 50% formamide, 5x SSC, and 1% SDS, incubating at 42°C, or, 5x SSC, 1 SDS, incubating at 65°C, with wash in 0.2x SSC, and 0.1% SDS at 65°C. Alternative conditions include, for example, conditions at least as stringent as hybridization at 68°C for 20 hours, followed by washing in 2x SSC, 0.1% SDS, twice for 30 minutes at 55°C and three times for 15 minutes at 60°C. Another alternative set of conditions is hybridization in 6x SSC at about 45°C, followed by one or more washes in 0.2x SSC, 0.1% SDS at 50-65°C. For PCR, a temperature

of about 36°C is typical for low stringency amplification, although annealing temperatures may vary between about 32°C and 48°C depending on primer length. For high stringency PCR amplification, a temperature of about 62°C is typical, although high stringency annealing temperatures can range from about 50°C to about 65°C, depending on the primer length and specificity. Typical cycle conditions for both high and low stringency amplifications include a denaturation phase of 90°C - 95°C for 30 sec. - 2 min., an annealing phase lasting 30 sec. - 2 min., and an extension phase of about 72°C for 1 - 2 min.

Nucleic acids that do not hybridize to each other under stringent conditions can be still substantially identical if they hybridize under moderately stringent conditions. Exemplary "moderately stringent hybridization conditions" include a hybridization in a buffer of 40% formamide, 1 M NaCl, 1% SDS at 37°C, and a wash in 1x SSC at 45°C. A positive hybridization is at least twice background. Those of ordinary skill will readily recognize that alternative hybridization and wash conditions can be utilized to provide conditions of similar stringency.

In certain embodiments, the invention includes fragments of functional sites. It is understood, as described above, that functional sites typically contain a core region associated with functional activity, as well as flanking regions. Accordingly, the invention includes fragments and regions of functional sites, including fragments consisting of or comprising core regions of functional sites. In certain embodiments, such fragments possess at least one physical or functional characteristic of the functional site from which they were derived. Functional fragments may be identified based upon any associated biological, biochemical, or physical function and by any available means. Thus, functional fragments of the invention include fragments capable of affecting or regulating (*e.g.* increasing or reducing) transcription of an operatively-linked gene, capable of binding to a transcription factor, capable of recruiting a transcriptional cofactor, capable of being methylated, and capable of directing methylation, demethylation,

acetylation, deacetylation, or any other modification of genomic DNA or chromatin, for example. Furthermore, it is not necessary that the functional fragment possesses the associated function in isolation; rather, a functional fragment may require the presence of additional regulatory or other nucleic acid sequences to function.

In one embodiment, a nucleic acid comprises between 10 and 75 bases identical to a sequence shown in Table 1. In another embodiment, a nucleic acid may comprise between 12 and 30, 15 to 50, 50 to 300, 100 to 200 or all of a sequence listed in Table 1. In most instances, at least 10 bases of a sequence desirably are used, preferably at least 20, and more preferably at least 50 bases. In additional embodiments, the present invention provides polynucleotide fragments comprising various lengths of contiguous stretches of sequence identical to or complementary to one or more of the sequences disclosed herein. For example, polynucleotides are provided by this invention that comprise at least about 10, 15, 20, 30, 40, 50, 75, 100, 150, 200, 300, 400, 500 or 1000 or more contiguous nucleotides of one or more of the sequences disclosed herein as well as all intermediate lengths there between. It will be readily understood that "intermediate lengths", in this context, means any length between the quoted values, such as 16, 17, 18, 19, *etc.*; 21, 22, 23, *etc.*; 30, 31, 32, *etc.*; 50, 51, 52, 53, *etc.*; 100, 101, 102, 103, *etc.*; 150, 151, 152, 153, *etc.*; including all integers through 200-500; 500-1,000, and the like.

In another embodiment, the invention includes fragments of functional site polynucleotides that do not possess a functional activity associated with the functional site. Such fragments may include, for example, probes or primers suitable for identifying, selecting or amplifying polynucleotides. Probes and primers of the invention include those corresponding to a region of a functional site or a complement thereof. In certain embodiments, probes and primers are preferably greater than 6 bases long, greater than 8, 10, 12, 16, or greater than 20 bases long. The term nucleic acid probe or oligonucleotide probe refers to a

nucleic acid capable of binding to a target nucleic acid of complementary sequence through one or more types of chemical bonds, usually through complementary base pairing and usually through hydrogen bond formation. As used herein, a probe includes natural *(i.e.,* A, G, C, or T) or modified bases (7-deazaguanosine, inosine, etc.). In addition, the bases in a probe may be joined by a linkage other than a phosphodiester bond, so long as it does not interfere with hybridization. It will be understood by one of skill in the art that probes may bind target sequences lacking complete complementarity with the probe sequence depending upon the stringency of the hybridization conditions. The probes may be directly labeled with isotopes, such as, for example, chromophores, lumiphores, or chromogens, or indirectly labeled, such as with biotin to which a streptavidin complex may later bind. The presence or absence of a target polynucletoide sequence of interest, such as a functional site, in a sample may be readily determined by determining the binding of a probe to the sample or the amplification of a PCR product from the sample.

In many embodiments, a functional site or nucleic acid of the invention is used at least in one stage as an isolated nucleic acid. The term isolated means a material that is at least partially free from components that normally accompany the material in the material's native state. Isolation connotes a degree of separation from an original source or surroundings. Isolated, as used herein, means that a polynucleotide is substantially away from other coding sequences, and that the DNA molecule does not contain large portions of unrelated coding DNA, such as large chromosomal fragments or other functional genes or polypeptide coding regions. Of course, this refers to the DNA molecule as originally isolated, and does not exclude genes or coding regions later added to the segment by the hand of man. By way of example and not limitation, a nucleic acid or peptide that is 0.1% pure in a biological sample becomes "isolated" when it is purified to at least 0.2% purity. In certain embodiments, the isolated material will become substantially free of cellular material, viral material, or culture medium

when produced by recombinant DNA techniques, or chemical precursors or other chemicals when chemically synthesized. Purity and homogeneity are typically determined using analytical chemistry techniques, for example, polyacrylamide gel electrophoresis or high performance liquid chromatography. An isolated DNA molecule prepared by chemical synthesis or enzymatic synthesis from cDNA represents another common example of isolated DNA. A skilled artisan knows a wide variety of procedures for preparing such isolated DNA via removing contaminants, thus making the DNA more homogeneous.

Nucleic acids that contain active genetic sequences may be of a variety of types, including deoxyribonucleotides or ribonucleotides and polymers thereof in either single- or double-stranded form. The term encompasses nucleic acids containing known nucleotide analogs or modified backbone residues or linkages, including synthetic, naturally occurring, and non-naturally occurring, which have similar binding properties as the reference nucleic acid. Examples of such analogs include, without limitation, phosphorothioates, phosphoramidates, methyl phosphonates, chiral methyl phosphonates, 2-O-methyl ribonucleotides, and peptide-nucleic acids (PNAs).

d. Vectors and Cells Comprising Functional Sites

Nucleic acids or polynucleotides of the invention may be inserted into vectors, including, for example, propagation and expression vectors, as described below. Vectors may include, but are not limited to, plasmids, episomes, baculovirus, retrovirus, lentivirus, adenovirus, and parvovirus, including adeno-associated virus. A variety of host cells may be used according to the invention, including, for example, mammalian cells, such as CHO, COS-7, or 293 cells. Other suitable host organisms include bacterial species (*e.g., E. coli* and *Bacillus*), other eukaryotes such as yeast (*e.g., Saccharomyces cerevisiae*), plant cells and insect cells (*e.g.*, Sf9). Suitable combinations of vectors and host cells are well

known in the art. Accordingly, vectors containing a polynucleotide of the invention and cells containing such a vector are among the aspects of the invention.

Vectors useful for the propagation of polynucleotide sequences are known and readily available in the art. Examples of such vectors include pUC vectors, pBluescript vectors, and pGEM vectors (Promega Corporation, Madison, WI). In certain embodiments, such vectors are capable of propagating in prokaryotic or eukaryotic cells, such as bacteria, *e.g.,* E. coli, or yeast, *e.g.,* S. cerevisiae.

Functional sites are useful in directing gene expression. In certain embodiments, the invention provides expression vectors comprising one or more polynucleotide sequences of the invention operably linked to a coding region, *e.g.,* such that the polynucleotides of the invention regulate expression of the coding region. Vectors comprising nucleic acids of the invention are also particularly useful in directing the expression of an associated or operably-linked gene in a specific cell type or developmental stage, since nucleic acid sequences of the invention include functional sites identified as being active in a specific cell type or under specific conditions. Such vectors may further comprise additional regulatory elements, such as, but not limited to, promoter sequences and enhancer sequences. A wide variety of suitable vectors for expression in eukaryotic cells are available. Such vectors include pCMVLacI, pXT1 (Stratagene Cloning Systems, La Jolla, CA); pCDNA series, pREP series, pEBVHis (Invitrogen, Carlsbad, CA).

Typical regulatory elements within vectors include a promoter sequence that contains elements that direct transcription of a linked gene and a transcription termination sequence. At minimum, the expression vector contains a promoter sequence, which may be the functional site sequence or a different promoter sequence. As used herein, a "promoter" refers to a nucleotide sequence that contains elements that direct the transcription of a linked gene. At minimum, a promoter contains an RNA polymerase binding site. When a promoter is linked to a gene so as to enable transcription of the gene, it is "operatively linked". The

promoter may be in the form of a promoter that is naturally associated with the gene of interest. Alternatively, the nucleic acid may be under control of a heterologous promoter not normally associated with the gene. Tissue specific promoter/enhancer elements may be used. In certain instances, the promoter elements may drive constitutive or inducible expression of the nucleic acid of interest. For expression in mammalian cells, mammalian promoters may be used, as well as viral promoters or any other promoter capable of driving expression in mammalian cells.

The expression vectors typically include a promoter designed for expression of the proteins in a desired host cell (*e.g.,* eukaryotic). Such promoters are widely available and are well known in the art. Promoters for expression in eukaryotic cells include the P10 or polyhedron gene promoter of baculovirus/insect cell expression systems (*see, e.g.*, U.S. Patent Nos. 5,243,041, 5,242,687, 5,266,317, 4,745,051, and 5,169,784), MMTV LTR, CMV IE promoter, RSV LTR, SV40, metallothionein promoter (*see, e.g.*, U.S. Patent No. 4,870,009) and the like.

Examples of other regulatory elements that may be present include secretion signal sequences, origins of replication, selectable markers, recombinase sequences, enhancer elements, nuclear localization sequences (NLS) and matrix association regions (MARS). In other preferred embodiments, the vector also includes a transcription terminator sequence. A "transcription terminator region" has either a sequence that provides a signal that terminates transcription by the polymerase that recognizes the selected promoter and/or a signal sequence for polyadenylation.

Preferably, vectors of the invention are capable of replication in the host cells. Thus, when the host cell is a bacterium, the vector preferably contains a bacterial origin of replication. Preferred bacterial origins of replication include the f1-ori and col E1 origins of replication, especially the ori derived from pUC plasmids. In yeast, ARS or CEN sequences can be used to assure replication. A well-used system in mammalian cells is SV40 ori.

The vectors also preferably include at least one selectable marker that is functional in the host. A selectable marker gene includes any gene that confers a phenotype on the host that allows transformed cells to be identified and selectively grown. Suitable selectable marker genes for bacterial hosts include the ampicillin resistance gene (Amp^r), tetracycline resistance gene (Tc^r) and the kanamycin resistance gene (Kan^r). The kanamycin resistance gene is presently preferred. Suitable markers for eukaryotes usually require a complementary deficiency in the host (*e.g.*, thymidine kinase (tk) in tk- hosts). However, drug markers are also available (*e.g.*, G418 resistance and hygromycin resistance).

A polynucleotide of the invention may further be coupled to a reporter gene to determine the regulatory activity of the polynucleotide. An example of a reporter gene is a nucleic acid that encodes an easily assayed protein such as chloramphenicol acetyltransferase. Examples of reporter genes include genes encoding alkaline phosphatase, β-galactosidase, and neomycin phosphotransferase. Other examples of reporter genes and their activities are shown in Table 3.

Table 3

Protein	Activity & Measurement
CAT (chloramphenicol acetyltransferase)	Transfers radioactive acetyl groups to chloramphenicol; detection by thin layer chromatography and autoradiography
GAL (β-galactosidase)	Hydrolyzes colorless galactosides to yield colored products.
GUS (β-glucuronidase)	Hydrolyzes colorless glucuronides to yield colored products.
LUC (luciferase)	Oxidizes luciferin, emitting photons.
GFP (green fluorescent protein)	Fluoresces on irradiation with UV.

Reporter genes may attach to other sequences so that only the reporter protein is made or so that the reporter protein is fused to another protein (fusion protein). Reporter genes "report" many different properties and events, for example: (i) the strength of promoters, whether native or modified for reverse genetics studies; (ii) the efficiency of gene delivery systems; (iii) the intracellular fate of a gene product, a result of protein traffic; (iv) the interaction of two proteins in the two-hybrid system or of a protein and a nucleic acid in the one-hybrid system; (v) the efficiency of translation initiation signals; and (vi) the success of molecular cloning efforts.

A host cell may be transformed with one or more vectors described herein. A host cell as termed herein means a naturally occurring cell or a transformed cell capable of supporting the replication or expression of the vector. Host cells may be cultured cells, explants, cells *in vivo*, and the like. Host cells may be prokaryotic cells, for example, *E. coli*, or eukaryotic cells, for example, yeast, insect, amphibian, or mammalian cells, for example, CHO, HeLa, and the like. Vectors may be introduced into host cells by a variety of methods well known in the art, depending upon the type of vector and corresponding host cell. Such methods are provided in detail in Molecular Cloning: A Laboratory Manual, Third Edition, eds. Sambrook *et al.* Cold Spring Harbor Press, 2001. For example, eukaryotic host cells may be transfected with plasmid or episomal vectors or infected with viral vectors. Bacterial or yeast host cells may be transformed with plasmid vectors, for example. Host cells may contain expression vectors in any manner, e.g., transiently, as episomes, or stably integrated into the host cell genome.

e. Transgenic, Knockout and Knockdown Reagents, Cells, Animals

The invention also includes transgenic and knock-out cells, plants, and animals comprising a disrupted nucleic acid sequence of the invention.

Transgenic and knockout cells of the invention include any suitable plant or animal, including humans and other mammals, such as mice, for example. Transgenic and knockout animals of the invention include suitable plants and non-human animals, including mice, or example. Methods for obtaining transgenic and knockout animals are known and available in the art.

i. Transgenic cells and animals

In one embodiment, the invention includes a transgenic animal that expresses a polynucleic acid or polypeptide, wherein expression is regulated by a nucleic acid of the invention. Accordingly, the invention further includes vectors suitable for the generation of a transgenic animal. Methods of generating transgenic animals are described, for example, in Hofker, M.H. (ed.), Van Deursen, J., and Sklar, H.T., (2002), TRANSGENIC MOUSE: METHODS AND PROTOCOLS (METHODS IN MOLECULAR BIOLOGY), Humana Press, Clifton, NJ. Transgenic cells and animals of the invention are particularly useful in providing or expressing a functional polypeptide in a particular cell or at a specific time in development or cell cycle, for example. A nucleic acid of the invention may be chosen to direct gene expression based upon the identification of the cell types and times during which it is active or hypersensitive.

ii. Knockout cells and animals

In another embodiment, a nucleic acid sequence of the invention is disrupted in an animal using knock-out methods, such that expression of a gene regulated by said sequence is altered. The invention, thus, includes knockout vectors, such as targeting or homologous recombination vectors, for example, and gene trap vectors, as well as cells an animals comprising all or part of such a vector within the genome of at some of their cells. Methods of generating a mouse containing an introduced gene disruption are described, for example, in Hogan, B.

et al., (1994), MANIPULATING THE MOUSE EMBRYO: A LABORATORY MANUAL, 2nd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

In one embodiment, gene targeting, which is a method of using homologous recombination to modify a cell's or animal's genome, can be used to introduce changes into cultured embryonic stem cells. By targeting a nucleic acid sequence of interest in ES cells, these changes can be introduced into the germlines of animals to generate chimeras and knock-out animals. Knockout cells and animals of the invention are useful in identifying genes regulated by the disrupted nucleic acid of the invention and the function of the disrupted nucleic acid of the invention.

Methods of designing and constructing vectors, and methods of introducing vector sequences into a genome are well known in the art and are described, for example, in GENE TARGETING: A PRACTICAL APPROACH, 2nd ed. (2000), Joyner, A.L., ed., Oxford University Press, New York; GENE TARGETING PROTOCOLS (METHODS IN MOLECULAR BIOLOGY, VOL. 133), (2000), Kmiec, E.B. and Gruenert, D.C., eds., Humana Press; and Torres, R.M. *et al.,* LABORATORY PROTOCOLS FOR CONDITIONAL GENE TARGETING (1997), Oxford University Press, Oxford; and references cited within, all of which are incorporated by reference. Specific vectors are also described in U.S. patents No. 5,364,783, No. 5,464,764 (positive-negative selection), No. 5,487,992, No. 5,627,059, No. 5,631,153, No. 5,719,055 (transposons), No. 5,830,698, No. 5,998, 144, No. 6,280,937, No. 6,284,541, No. 6,139,833, 6,303,327, No. 6,319,692, No. 6,329,200, and No. 6,080,576, and references and patents cited therein. Generally, vectors of the invention may be constructed, propagated, isolated, and examined using routine molecular biology techniques such as restriction enzyme digestion, polymerase chain reaction, ligation, transformation, and southern blotting, according to procedures well known in the art and described in CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, (2001), Ausubel *et al.* (eds.), John Wiley & Sons, New York and Sambrook, *et al.*, MOLECULAR CLONING: A LABORATORY MANUAL, (2001), Cold

Spring Harbor Laboratory Press, Cold Spring Harbor, New York, and U.S. Patent No. 5,789,215, for example.

iii. Knockdown cells and animals

In certain embodiments, the invention includes knockdown reagents targeted to functional sites or genes associated with or regulated by functional sites, and cells and animals comprising said reagents. Such knockdown reagents may be used, for example, to alter, *e.g.* reduce, increase, or disrupt, the expression of one or more genes regulated by a targeted functional site. Knockdown reagents include any of a variety of agents that may reduce mRNA levels. Knockdown reagents include, for example, ribozymes, antisense RNA, and double-stranded RNAs, including small interfering RNAs (siRNAs) and short hairpin RNAs (shRNAs).

Antisense oligonucleotides have been demonstrated to be effective and targeted inhibitors of protein synthesis, and, consequently, can be used to specifically inhibit protein synthesis by a targeted gene. The efficacy of antisense oligonucleotides for inhibiting protein synthesis is well established. For example, the synthesis of polygalactauronase and the muscarine type 2 acetylcholine receptor are inhibited by antisense oligonucleotides directed to their respective mRNA sequences (U. S. Patent 5,739,119 and U. S. Patent 5,759,829). Further, examples of antisense inhibition have been demonstrated with the nuclear protein cyclin, the multiple drug resistance gene (MDG1), ICAM-1, E-selectin, STK-1, striatal $GABA_A$ receptor and human EGF (Jaskulski *et al.*, Science. 1988 Jun 10;240(4858):1544-6; Vasanthakumar and Ahmed, Cancer Commun. 1989;1(4):225-32; Peris *et al.*, Brain Res Mol Brain Res. 1998 Jun 15;57(2):310-20; U. S. Patent 5,801,154; U.S. Patent 5,789,573; U. S. Patent 5,718,709 and U.S. Patent 5,610,288). Furthermore, antisense constructs have also been described that inhibit and can be used to treat a variety of abnormal cellular

proliferations, *e.g.* cancer (U. S. Patent 5,747,470; U. S. Patent 5,591,317 and U. S. Patent 5,783,683).

Therefore, in certain embodiments, the present invention provides oligonucleotide sequences that comprise all, or a portion of, any sequence that is capable of specifically binding to a selected target polynucleotide sequence, or a complement thereof. In one embodiment, the antisense oligonucleotides comprise DNA or derivatives thereof. In another embodiment, the oligonucleotides comprise RNA or derivatives thereof. The antisense oligonucleotides may be modified DNAs comprising a phosphorothioated modified backbone. Also, the oligonucleotide sequences may comprise peptide nucleic acids or derivatives thereof. In each case, preferred compositions comprise a sequence region that is complementary, and more preferably, completely complementary to one or more portions of a target gene or polynucleotide sequence. Selection of antisense compositions specific for a given sequence is based upon analysis of the chosen target sequence and determination of secondary structure, T_m, binding energy, and relative stability. Antisense compositions may be selected based upon their relative inability to form dimers, hairpins, or other secondary structures that would reduce or prohibit specific binding to the target mRNA in a host cell. Highly preferred target regions of the mRNA include those regions at or near the AUG translation initiation codon and those sequences that are substantially complementary to 5' regions of the mRNA. These secondary structure analyses and target site selection considerations can be performed, for example, using v.4 of the OLIGO primer analysis software and/or the BLASTN 2.0.5 algorithm software (Altschul *et al.,* Nucleic Acids Res. 1997, 25(17):3389-402).

The use of an antisense delivery method employing a short peptide vector, termed MPG (27 residues), is also contemplated. The MPG peptide contains a hydrophobic domain derived from the fusion sequence of HIV gp41 and a hydrophilic domain from the nuclear localization sequence of SV40 T-antigen (Morris *et al.*, Nucleic Acids Res. 1997 Jul 15;25(14):2730-6). It has been

demonstrated that several molecules of the MPG peptide coat the antisense oligonucleotides and can be delivered into cultured mammalian cells in less than 1 hour with relatively high efficiency (90%). Further, the interaction with MPG strongly increases both the stability of the oligonucleotide to nuclease and the ability to cross the plasma membrane.

According to another embodiment of the invention, ribozyme molecules are used to inhibit expression of a target gene or polynucleotide sequence. Ribozymes are RNA-protein complexes that cleave nucleic acids in a site-specific fashion. Ribozymes have specific catalytic domains that possess endonuclease activity (Kim and Cech, Proc Natl Acad Sci U S A. 1987 Dec;84(24):8788-92; Forster and Symons, Cell. 1987 Apr 24;49(2):211-20). For example, a large number of ribozymes accelerate phosphoester transfer reactions with a high degree of specificity, often cleaving only one of several phosphoesters in an oligonucleotide substrate (Cech *et al.*, Cell. 1981 Dec;27(3 Pt 2):487-96; Michel and Westhof, J Mol Biol. 1990 Dec 5;216(3):585-610; Reinhold-Hurek and Shub, Nature. 1992 May 14;357(650):173-6). This specificity has been attributed to the requirement that the substrate bind via specific base-pairing interactions to the internal guide sequence ("IGS") of the ribozyme prior to chemical reaction.

At least six basic varieties of naturally-occurring enzymatic RNAs are known presently. Each can catalyze the hydrolysis of RNA phosphodiester bonds *in trans* (and thus can cleave other RNA molecules) under physiological conditions. In general, enzymatic nucleic acids act by first binding to a target RNA. Such binding occurs through the target binding portion of an enzymatic nucleic acid that is held in close proximity to an enzymatic portion of the molecule that acts to cleave the target RNA. Thus, the enzymatic nucleic acid first recognizes and then binds a target RNA through complementary base-pairing, and once bound to the correct site, acts enzymatically to cut the target RNA. Strategic cleavage of such a target RNA will destroy its ability to direct synthesis of an encoded protein. After an enzymatic nucleic acid has bound and cleaved its RNA target, it is

released from that RNA to search for another target and can repeatedly bind and cleave new targets.

The enzymatic nature of a ribozyme may be advantageous over many technologies, such as antisense technology (where a nucleic acid molecule simply binds to a nucleic acid target to block its translation), since the concentration of ribozyme necessary to affect inhibition of expression is lower than that of an antisense oligonucleotide. This advantage reflects the ability of the ribozyme to act enzymatically. Thus, a single ribozyme molecule is able to cleave many molecules of target RNA. In addition, the ribozyme is a highly specific inhibitor, with the specificity of inhibition depending not only on the base pairing mechanism of binding to the target RNA, but also on the mechanism of target RNA cleavage. Single mismatches, or base-substitutions, near the site of cleavage can completely eliminate catalytic activity of a ribozyme. Similar mismatches in antisense molecules do not prevent their action (Woolf *et al.*, Proc Natl Acad Sci U S A. 1992 Aug 15;89(16):7305-9). Thus, the specificity of action of a ribozyme is greater than that of an antisense oligonucleotide binding the same RNA site.

The enzymatic nucleic acid molecule may be formed in a hammerhead, hairpin, a hepatitis δ virus, group I intron or RNaseP RNA (in association with an RNA guide sequence) or Neurospora VS RNA motif, for example. Specific examples of hammerhead motifs are described by Rossi *et al.* Nucleic Acids Res. 1992 Sep 11;20(17):4559-65. Examples of hairpin motifs are described by Hampel *et al.* (Eur. Pat. Appl. Publ. No. EP 0360257), Hampel and Tritz, Biochemistry 1989 Jun 13;28(12):4929-33; Hampel *et al.*, Nucleic Acids Res. 1990 Jan 25;18(2):299-304 and U. S. Patent 5,631,359. An example of the hepatitis δ virus motif is described by Perrotta and Been, Biochemistry. 1992 Dec 1;31(47):11843-52; an example of the RNaseP motif is described by Guerrier-Takada *et al.*, Cell. 1983 Dec;35(3 Pt 2):849-57; Neurospora VS RNA ribozyme motif is described by Collins (Saville and Collins, Cell. 1990 May 18;61(4):685-96; Saville and Collins, Proc Natl Acad Sci U S A. 1991 Oct 1;88(19):8826-30; Collins

and Olive, Biochemistry. 1993 Mar 23;32(11):2795-9); and an example of the Group I intron is described in (U. S. Patent 4,987,071). Important characteristics of enzymatic nucleic acid molecules used according to the invention are that they have a specific substrate binding site which is complementary to one or more of the target gene DNA or RNA regions, and that they have nucleotide sequences within or surrounding that substrate binding site which impart an RNA cleaving activity to the molecule. Thus the ribozyme constructs need not be limited to specific motifs mentioned herein.

Ribozymes may be designed as described in Int. Pat. Appl. Publ. No. WO 93/23569 and Int. Pat. Appl. Publ. No. WO 94/02595, each specifically incorporated herein by reference, and synthesized to be tested *in vitro* and *in vivo*, as described. Such ribozymes can also be optimized for delivery. While specific examples are provided, those in the art will recognize that equivalent RNA targets in other species can be utilized when necessary.

Ribozyme activity can be optimized by altering the length of the ribozyme binding arms, or chemically synthesizing ribozymes with modifications that prevent their degradation by serum ribonucleases (see *e.g.*, Int. Pat. Appl. Publ. No. WO 92/07065; Int. Pat. Appl. Publ. No. WO 93/15187; Int. Pat. Appl. Publ. No. WO 91/03162; Eur. Pat. Appl. Publ. No. 92110298.4; U. S. Patent 5,334,711; and Int. Pat. Appl. Publ. No. WO 94/13688, which describe various chemical modifications that can be made to the sugar moieties of enzymatic RNA molecules), modifications which enhance their efficacy in cells, and removal of stem II bases to shorten RNA synthesis times and reduce chemical requirements.

RNA interference methods using double-stranded RNA also may be used to disrupt the expression of a gene or polynucleotide of interest. A dsRNA molecule that targets and induces degradation of an mRNA that is derived from a gene or polynucleotide of interest can be introduced into a cell. The exact mechanism of how the dsRNA targets the mRNA is not essential to the operation of the invention, other than the dsRNA shares sequence homology with the mRNA

transcript. The mechanism could be a direct interaction with the target gene, an interaction with the resulting mRNA transcript, an interaction with the resulting protein product, or another mechanism. Again, while the exact mechanism is not essential to the invention, it is believed the association of the dsRNA to the target gene is defined by the homology between the dsRNA and the actual and/or predicted mRNA transcript. It is believed that this association will affect the ability of the dsRNA to disrupt the target gene. DsRNA methods and reagents are described in PCT application WO 01/68836, WO 01/29058, WO 02/44321, and WO 01/75164, which are hereby incorporated by reference in their entirety.

In one embodiment of the invention, double-stranded RNA interference (dsRNAi) may be used to specifically inhibit target nucleic acid expression. Briefly, it is hypothesized that the presence of double-stranded RNA dominantly silences gene expression in a sequence-specific manner by causing the corresponding RNA to be degraded. Although first discovered in lower organisms such as the nematode and *Drosphila*, for example, dsRNAi has also been demonstrated to work in fungi, plants, and mammalian cells (Wianny, F. and Zernica-Goetz, M. (2000), Nature Cell Biology Vol. 2, 70-75). However, transfection of long dsRNAs into mammalian cells can result in nonspecific gene suppression, as opposed to the gene-specific suppression observed in other organisms.

Although the mechanisms behind dsRNAi is still not entirely understood, experiments demonstrated that, in the cell, a double-stranded RNA (dsRNA) is cleaved into short pieces, typically 21-25 nucleotides in length, termed small interfering RNAs (siRNAs), by a ribonuclease such as DICER. The siRNAs subsequently assemble with protein components into an RNA-induced silencing complex (RISC), which binds to and tags the complementary portion of the target mRNA for nuclease digestion. The siRNA triggers the degradation of mRNA that matches its sequence, thereby repressing expression of the corresponding gene.

Discussed in Bass, B. Nature 411:428-429 (2001) and Sharp, P.A. Genes Dev. 15:485-490 (2001).

Double-stranded RNA-mediated suppression of gene and nucleic acid expression may be accomplished according to the invention by introducing dsRNA, siRNA or shRNA into cells or organisms. dsRNAs less than 30 nucleotides in length do not appear to induce nonspecific gene suppression, as described above for long dsRNA molecules. Indeed, the direct introduction of siRNAs to a cell can trigger RNAi in mammalian cells (Elshabir, S.M., *et al.* Nature 411:494-498 (2001)). Furthermore, suppression in mammalian cells occurred at the RNA level and was specific for the targeted genes, with a strong correlation between RNA and protein suppression (Caplen, N. *et al.*, Proc. Natl. Acad. Sci. USA 98:9746-9747 (2001)). In addition, it was shown that a wide variety of cell lines, including HeLa S3, COS7, 293, NIH/3T3, A549, HT-29, CHO-KI and MCF-7 cells, are susceptible to some level of siRNA silencing (Brown, D. *et al.* TechNotes 9(1):1-7, available at http://www.ambion.com/techlib/tn/91/912.html (9/1/02)).

Structural characteristics of effective siRNA molecules have been identified. Elshabir, S.M. *et al.* (2001) Nature 411:494-498 and Elshabir, S.M. *et al.* (2001), EMBO 20:6877-6888. Accordingly, one of skill in the art would understand that a wide variety of different siRNA molecules may be used to target a specific gene or transcript. In certain embodiments, siRNA molecules according to the invention are 18 - 25 nucleotides in length, including each integer in between. In one embodiment, an siRNA is 21 nucleotides in length. In certain embodiments, siRNAs have 0-7 nucleotide 3' overhangs or 0-4 nucleotide 5' overhangs. In one embodiment, an siRNA molecule has a two nucleotide 3' overhang. In one embodiment, an siRNA is 21 nucleotides in length with two nucleotide 3' overhangs (*i.e.* they contain a 19 nucleotide complementary region between the sense and antisense strands). In certain embodiments, the overhangs are UU or dTdT 3' overhangs. Generally, siRNA molecules are completely complementary to one strand of a target DNA molecule, since even

single base pair mismatches have been shown to reduce silencing. In other embodiments, siRNAs may have a modified backbone composition, such as, for example, 2'-deoxy- or 2'-O-methyl modifications. However, in preferred embodiments, the entire strand of the siRNA is not made with either 2' deoxy or 2'-O-modified bases.

In one embodiment, siRNA target sites are selected by scanning the target mRNA transcript sequence for the occurrence of AA dinucleotide sequences. Each AA dinucleotide sequence in combination with the 3' adjacent approximately 19 nucleotides are potential siRNA target sites. In one embodiment, siRNA target sites are preferentially not located within the 5' and 3' untranslated regions (UTRs) or regions near the start codon (within approximately 75 bases), since proteins that bind regulatory regions may interfere with the binding of the siRNP endonuclease complex (Elshabir, S. et al. Nature 411:494-498 (2001); Elshabir, S. et al. EMBO J. 20:6877-6888 (2001)). In addition, potential target sites may be compared to an appropriate genome database, such as BLAST, available on the NCBI server at www.ncbi.nlm, and potential target sequences with significant homology to other coding sequences eliminated.

Short hairpin RNAs may also be used to inhibit or knockdown gene or nucleic acid expression according to the invention. Short Hairpin RNA (shRNA) is a form of hairpin RNA capable of sequence-specifically reducing expression of a target gene. Short hairpin RNAs may offer an advantage over siRNAs in suppressing gene expression, as they are generally more stable and less susceptible to degradation in the cellular environment. It has been established that such short hairpin RNA-mediated gene silencing (also termed SHAGging) works in a variety of normal and cancer cell lines, and in mammalian cells, including mouse and human cells. Paddison, P. *et al.*, Genes Dev. 16(8):948-58 (2002). Furthermore, transgenic cell lines bearing chromosomal genes that code for engineered shRNAs have been generated. These cells are able to constitutively synthesize shRNAs, thereby facilitating long-lasting or constitutive gene silencing

that may be passed on to progeny cells. Paddison, P. *et al.*, Proc. Natl. Acad. Sci. USA 99(3):1443-1448 (2002).

ShRNAs contain a stem loop structure. In certain embodiments, they may contain variable stem lengths, typically from 19 to 29 nucleotides in length, or any number in between. In certain embodiments, hairpins contain 19 to 21 nucleotide stems, while in other embodiments, hairpins contain 27 to 29 nucleotide stems. In certain embodiments, loop size is between 4 to 23 nucleotides in length, although the loop size may be larger than 23 nucleotides without significantly affecting silencing activity. ShRNA molecules may contain mismatches, for example G-U mismatches between the two strands of the shRNA stem without decreasing potency. In fact, in certain embodiments, shRNAs are designed to include one or several G-U pairings in the hairpin stem to stabilize hairpins during propagation in bacteria, for example. However, complementarity between the portion of the stem that binds to the target mRNA (antisense strand) and the mRNA is typically required, and even a single base pair mismatch is this region may abolish silencing. 5' and 3' overhangs are not required, since they do not appear to be critical for shRNA function, although they may be present (Paddison *et al.* (2002) Genes & Dev. 16(8):948-58).

f. Libraries and Arrays Comprising Functional Site Sequences

The invention further includes libraries and arrays of polynucleotide sequences comprising functional sites or fragments, complements or variants thereof. Libraries of functional sites are useful for a variety of purposes set forth herein, including, for example, identifying sequences that coordinately regulate specific genes or sets of genes. A library comprises at least two polynucleotide sequences of the invention or fragments or functional fragments thereof. Libraries may comprise isolated nucleic acid fragments, vectors comprising inserts corresponding to nucleic acid sequences of the invention, or cells comprising such

vectors, for example. A library or "set" as termed here may have at least two members selected from Table 1 and may have at least 10 members, 100 members, 500 members, 1,000 members, 2,000 members, 5,000 members, 10,000 members, 20,000 members or even more than 30,000 members selected from these sequences.

Libraries of the invention may include functional sites located throughout the genome, *i.e.,* genome-wide libraries, or they may include functional sites associated with a specific region of the genome, such as, for example, a particular chromosome or within a particular distance of a known gene or mutation. In one embodiment, the invention provides a set of members of functional sites sites associated with chromosome 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,18, 19, 20, 21, 22, or 23.

Libraries and arrays may also include functional sites associated with any recognizable of definable group, such as those described *supra*. Thus, a library may, for example, include functional sites associated with a particular cell or tissue type. These functional sites may be unique or specific to the particular cell or tissue, or they may be present in a number of tissues or constitutive, *i.e.* active in most or all cells and tissues. In one exemplary embodiment, the invention includes a library or array of functional sites associated with a particular cell type following the administration of an agent, such as a drug. In another embodiment, a library includes two or more functional sites or mutant functional sites associated with a specific disease or disorder, such as a neoplasia. In certain embodiments, a library or array comprises to or more functional sites or functional site mutants or variants associated with a leukemia (*e.g.,* chronic myelogenous leukemia an acute myelogenous leukemia), hepatic carcinoma, breast cancer, prostate cancer, or lung cancer. Additional examples diseases or disorders with which the functional sites are associated are listed in Column 12 of Table 1.

The sequences further may be used as substrates for producing cross-referenced libraries to define key active genetic elements. Many

hypersensitive sites are common between tissues and cells exposed to different stimuli. For example, some hypersensitive sites are associated with constitutively expressed genes, and some are unique and define the cell and its transcriptional program. To find these differentially formed hypersensitive sites, subtracted libraries can be made using hypersensitive sites cloned from two different populations as substrates.

The present invention also provides arrays, including microarrays, of nucleic acids of the invention or fragments or functional fragments thereof. In addition, arrays may comprise cells of the invention. In certain embodiments, such arrays comprise two or more different polynucleotides or cells, each located at a discrete and positional addressable or identifiable position on a solid support or in discrete vessels. In another embodiment, an array may comprise a plurality of different polynucleotides or cells, with several different polynucleotides or cells located at a discrete position on a solid support or in a discrete vessel. Preferably, each position or vessel comprises 1, between 1 and 3, between 3 and 10, or between 10 and 100 different polynucleotides or cells. These polynucleotides or cells may be located in a positionally addressable location on the array, such that the identity of functional sites located at each location on the array is known or may be readily determined. Methods and procedures for producing arrays are well known in the art and include those described, for example, in U.S. Patent Application Serial Nos. 10/319,440, filed December 12, 2002, and 10/375,404, filed February 27, 2003, and PCT Application No. PCT/US02/15032, filed May 12, 2002, and references cited therein, each of which is hereby incorporated by reference in its entirety.

In particular embodiments, the invention provides arrays of polynucleotides comprising functional site sequences set forth in Table 1 or fragments or complements thereof. Such arrays may comprise or consist of sets of functional sites sequences, including, for example, sets of functional site sequences associated with a particular chromosome, gene, or other genomic

locus, sets of functional site sequences associated with a disease, clinical outcome, or therapeutic response.

B. Methods of Identifying Functional Sites

A variety of methods may be employed to identify functional site sequences of the invention. Detailed descriptions of methods of identifying and isolating functional sites are provided in U.S. Provisional Patent Applications No. 60/108,206, No. 60/302,369, and No. 60/290,036, U.S. Patent Applications Serial No. 09/432,576, Serial No. 10/187,887, Serial No. 10/157,027, and Serial No. 10/319,440, PCT Publication No. WO 02/097135, and PCT Application No. PCT/US02/15032, which are hereby incorporated by reference in their entirety. In addition, polynucleotides may be cloned from genomic libraries by routine procedures, including, or example, polymerase chain reaction, or synthesized using techniques well known in the art.

In one embodiment, a general method of identifying functional sites includes the basic steps of: (1) treating nuclear chromatin with an agent that cleaves or tags DNA at functional sites; and (2) isolating DNA segments flanking cleavage sites or tagged sites. In addition, the isolated DNA segments may be subcloned into a vector. The basic method may also be performed using *in vitro* assembled chromatin constructs. In one embodiment, the method further includes the step of amplifying the isolated DNA segments before subcloning, preferably by PCR.

A variety of agents may be used to cleave or tag functional sites. Any agent capable of detecting a focal alteration in chromatin structure may be employed to identify functional site sequences. Functional sites are modified by the action of one or more of these factors on the biological sample, the best documented and recognized example of which is the action of the non-specific endonuclease DNAse (*e.g.* EMBO J 14:106-16 (1995)). Non-specific endonucleases, such as Dnasel, are typically used to discover functional sites, but

other agents can be used just as well. Potentially a subset of functional sites will note be detected by DNAse I and sets of functional sites may alternatively be identified by the actions of nucleases (both sequence-specific and non-specific), endogenous and exogenous); topoisomerases; methylases; acetylases; chemicals; pharmaceuticals (*e.g.* chemotherapy agents); radiation; physical shearing; nutrient deprivation (*e.g.* folate deprivation); etc. Essentially any agent, whether biological (*e.g.* enzymes), chemical (*e.g.* DNA bidning molecules), or physical (*e.g.* stress), which will modify DNA in the nucleus, which is not occluded in the folded chromatin structure but exists in open regions accessible to DNA binding activities and is, hence, more liable to break. For example, modifications of the DNA in the nucleus, such as the action of dam methylase, can be used as a marker when the DNA is subsequently purified, for example, by the use of restriction enzymes that are differentially sensitive to dam methylation. Exemplary classes of these agents and examples of such are set forth in Table 4.

Table 4. Agents Suitable for Detection of Functional Sites

Class	*Description*	*Example*	*Site examined*	*Reference*
Non-specific nucleases	Endonucleases with little or no cutting specificity	DNaseI, DNaseII, Micrococcal nuclease	Chicken □□ globin 5' HS1	Wood and Felsenfeld, 1982
Endogenous nucleases		DNaseI		
Restriction endonucleases	Sequence-specific endonucleases	*Pvu* II, *Nhe* I	Chicken erythroid-specific □^{A}/□-globin enhancer	Boyes and Felsenfeld, 1997
Modified DNA-binding proteins	Synthetic proteins capable of binding within sites of interest and inducing cutting or modification	Sp1 + nuclease tail (PIN*POINT)	Human MnSOD promoter	Kuo *et al.*, 2002
DNA modifying enzymes	DNA-binding enzymes which modify their binding site	*dam* DNA methyltransferase	*lacZ* reporter gene in *Drsophila* nuclei	Wines *et al.*, 1996

Intercalator agents	DNA minor and major groove intercalators that cause strand breakage	Bleomycin		
Topoisomerases	Naturally-occurring nuclear enzymes that change DNA linking number via single- or double-strand breakage, DNA strand rotation, and re-ligation	Topo II		
Viruses	Viruses that integrate into the genome			

Alternatively, specific classes of functional sites may be targeted. For example, those known to be bound by a specific protein can be enriched for either by adding exogenous modified protein, which binds to its recognition site with in the functional site and induces modification (*e.g.* by creating a chimeric DNA-binding protein with a methylase or by incorporation of cross-linking reagents such as 4-azidophenacylbromide (e.g. Proc. Natl. Acad. Sci USA 89: 10287-10291) or strand damage (e.g. by incorporation of 125I, the radioactive decay of which would cause strand breakage (e.g. Acta Oncol. 39: 681-785 (2000)). Advantag can also be taken of such proteins bound in their natural context by isolating the nucleoprotein complexes in chromatin containing such proteins via antibody recogniztion (the ChIp protocol, Orlando et al., Methods 11:205-214 (1997)).

An alternate approach is to produce functional site enriched samples by fractionation. Digestion of nuclei will create a population of fragments where the smaller ones are more likely to have one or more cut sites within functional sites. That is as, dependent on the digestion conditions, wither a functional site has received more than one cut to produce a small fragment whereas the background remains large. Alternatively, the functional site has been cut once, but the average distance between a functional site-cut and random cut or shear site is smaller than the average size of the entire population. Fragments can be

separated on the basis of their size, before or after purification of the DNA from chromatin, by various methods including ultracentrifugation, preparative gel electrophoresis or size exclusion columns. If the fragments are isolated from the nuclei as chromatin fractions, they can be further enriched for functional site-containing material prior to centrifugation on the basis of properties of the nucleoprotein complexes that distinguish them from bulk chromatin. These include, for example, higher salt solubility of active chromatin domains (Ridsdale *et al.* Nucl. Acids. Res. 16:5915-5926 (1988)), the reactivity of thiol groups on the histone H3 (Chen-Cleland *et al.,* J. Biol. Chem. 268:23409-23416 (1993)) and the extraction of nucleosomal DNA by binding to sulfated polysaccharides, such as heparin (Watson *et al.*, J. Biol. Chem. 274:21707-21703).

Similarly, a variety of different methods may be utilized to isolate DNA segments containing functional sites, including the use of linkers, streptavidin/biotin, magnetic beads, and ab/hapten systems, for example. In certain embodiments, isolated functional sites may be labeled, *e.g.* when used to probe an array. The labeling of functional sites is achieved by standard methods, *e.g.,* performing amplifications (linear or exponential) using synthetically labeled oligonucleotides (*e.g.* containing Cy5- or Cy3-modified nucleotides or amino allyl modified nucleotides, which allow for chemical coupling of dye molecules post-amplification), or by direct incorporation of modified nucleotides during the reaction.

Additional embodiments of the general method of identifying functional sites include using subtractive methods designed to enrich functional site sequences and/or identify cell-specific functional sites. Subtractive methods may also be employed to remove repetitive sequences.

Another embodiment of the method of identifying functional sites involves concatamerizing isolated DNA segments, typically after further digesting the isolated fragments with a type IIs restriction enzyme to generate fragments of uniform size. The concatamer approach permits the sequencing and identification

of multiple functional sites within a single polynucleotide sequence. In certain embodiment, linker sequences may be attached to one or more ends of the isolated fragments prior to concatamerization, typically by ligation. The boundaries of each isolated DNA segment, comprising a functional site, is readily determined by identifying the restriction site sequence or linker sequence located at one or both ends of each isolated DNA segment within the polynucleotide produced upon concatamerization.

In one embodiment, the sensitivity of a region of genomic DNA to DNA-modifying agents is quantified using Real-Time PCR. Such methods allow quantitative characterization of the activity of functional sites and the identification of functional sites with cell-specific or disrupted activities. The method generally involves isolating chromatin, treating a portion of the chromatin with a DNA modifying agent, treating another portion of the chromatin with the DNA modifying agent under modified conditions, isolating treated DNA from each portion, amplifying the candidate region by Real-Time PCR from each portion, determining copy number of the candidate region, and comparing to a reference curve to obtain relative copy number ratio of the candidate region and the reference region. Thus, the sensitivity of the candidate region to the DNA modifying agent is thereby determined relative to the sensitivity of the reference region. Embodiments of this method may also be used to detect single stranded nicks and to quantify naturally occurring single stranded DNA structures *in vivo*.

C. Methods of Using Polynucleotides Comprising Functional Sites

The functional site sequences and genome location information provided in Table 1 may be used in whole or in part, individually or linked in combination with other sequences, as discovery tools and medical agents. The functional site sequences of the invention, alone or in combination with their genomic locations, may be used in a variety of different ways to identify genes and cells, including, *e.g.,* disease genes and cells, and regulate gene expression. For

example, functional site sequences may be used to make new DNA fragments, vectors, transgenic cells, transgenic organisms, and for new diagnostics that allow better diagnosis and intervention in disease affected by gene regulation.

Since functional site sequences are regulatory elements involved in directing gene expression, functional sites of the invention may be employed according to the invention to direct gene expression. Furthermore, since the invention provides functional sites associated with specific cells or tissues, specific functional sites may be used to direct expression to a particular cell type. Functional site sequences may be physically linked to a gene of interest, *e.g.,* in an expression vector, in order to direct expression of the gene, *e.g.,* to a particular cell type or during a specific stage of differentiation or development. Accordingly, it is understood that functional sites may be used to direct gene expression in any context, including, but not limited to, gene therapy or replacement therapy and transgenic expression.

Furthermore, functional site activity provides a unique means of characterizing and identifying cells, which has broad diagnostic and therapeutic applications. Gene expression differs between different cell types, between normal and diseased cells, and in response to different stimuli. The differences in gene expression are mediated, in large part, by differing activity of regulatory sequences, *i.e.* functional sites, in different cells. Thus, different cell types have different functional sites. Accordingly, functional site activity may be used to characterize cells. The skilled artisan would immediately recognize that methods of characterizing and identifying cells by examining the functional sites present within the cell have broad applications and may be used to identify cells for any purpose. Examples include, but are not limited to, tissue-typing, determining the primary source of metastatic cells, diagnosing a disease, determining whether a cell has been exposed to a specific stimuli or drug, determining the growth or differentiation state of a cell, and determining the developmental stage of a cell.

Functional sites, or fragments thereof, present in a particular cell or sample may be isolated according to the methods described herein, and the presence of specific sites within these isolated DNA sequences may be determined by a variety of means, including, *e.g.,* subcloning and sequencing of the isolated DNA sequences, hybridization to a panel or array of functional site sequences, direct sequencing, and PCR amplification using primers specific for functional site sequences. Functional sites specific to a cell or treatment may be identified by any available means, including, *e.g.,* comparing functional site sequences from different cells, subtractive hybridization methods, and screening an array of functional sites and determining which functional sites in the array are present in the sample. Methods of screening arrays of functional site sequences are provided in detail in U.S. Patent Application Serial No. 10/319,440, U.S. Patent Application Serial No. 10/375,404, and PCT Patent Application No. PCT/US02/15032, which are hereby incorporated by reference in their entirety. Functional sites specific for a cell or treatment may be identified by any available means in the art, as would readily be understood by the skilled artisan, and include, *e.g.*, comparison of functional site sequences identified in different cells, subtractive hybridization methods, and probing an array of functional site sequences.

A variety of methods employing the sequences and information provided by the invention are set forth in further detail below, by means of illustration but not limitation.

1. Regulation of Gene Expression

Nucleic acids and sequences of the invention may be used to regulate the expression of a gene. Accordingly, the invention provides methods of regulating gene expression using one or more polynucleotide sequences of the invention. These methods include, but are not limited to, expression of a heterologous gene, gene therapy methods, transgenic methods, and knockout and

knockdown methods. Such methods may be employed to direct expression of a gene or nucleic acid in a particular cell or tissue type, for example, by placing the gene under the control of one or more functional sites identified as present or specific to the desired tissue or cell type. In addition, such methods may be employed to increase or reduce expression of a gene or nucleic acid. It is understood that a gene or nucleic acid being regulated in any o these manner includes both protein encoding DNA sequences and non-protein encoding sequences, *e.g.,* sequences encoding a functional RNA molecule, such as an antisense RNA or shRNA. Furthermore, genes or nucleic acids may be regulated according to the invention *in vitro* or *in vivo*, and the expression, knockout, or knockdown constructs may be either transiently present within a cell or stably integrated into a cellular genome.

In certain embodiments, expression of endogenous genes may be regulated, for example, by introducing a sequence of the invention into the genome of a cell, plant or animal, for example, by homologous recombination or gene trap methods. Alternatively, expression of an endogenous gene may be regulated using knockdown reagents, such as antisense, directed to or complementary to a sequence of the invention or a gene regulated by a sequence of the invention. In certain embodiment, such knockdown methods target functional sites located within transcribed regions of genomic DNA. In other embodiments, an exogenous gene or polynucleotide sequence regulated by a sequence of the invention may be introduced into a cell or animal. As described in more detail below, the use of functional sites in expression vectors permits targeted expression, *e.g.,* to a particular cell or tissue type or during a particular stage of differentiation or cell cycle. Of course, it is understood that such methods may employ one or a plurality of functional sites. When multiple functional sites are used, they may all be different, or they may comprise more than one copy of one or more functional sites.

In one embodiment of regulating an endogenous gene by introducing a functional site sequence into the genome, the sequence is inserted at a position relative to the gene to be regulated based upon the genomic position of the sequence relative to a gene it normally regulates, *e.g.,* upstream or downstream and the approximate distance from the transcribed region of the gene, although the skilled artisan would understand that a variety of functional site sequences are active at multiple positions in relation to a regulated gene, so the sequence may be inserted at any position relative to the gene desired to be regulated. In particular embodiments, functional site sequences may be inserted into a genome within at least 10, 20, 50, 100, 500, 1000, or 5000 base pairs of the transcription start site or promoter of a gene to be regulated. In certain embodiments, nucleic acid sequences may be inserted to restore normal expression levels or patterns of a gene, *e.g.,* gene therapy, or to alter expression of a gene, for example. Thus, in one embodiment, targeting methods may be used to functionally or physically replace a mutated or defective functional site.

The sequences advantageously may be used as substrates for recombination. One of the strategies for making a causal link between a functional site and its function is to carry out a recombination experiment. For example, one can conditionally knock out active genetic elements and monitor changes in phenotype. This strategy utilizes the sequence of a cloned functional sites to design constructs that can cause recombination and loss, or conditional loss, of the functional site sequence *in vivo*. Accordingly, the invention includes methods of knocking out or mutating a functional site sequence, involving a homologous recombination or targeting vector comprising a functional site sequence.

The regulation of a gene may be altered by knocking out, *e.g.,* via homologous recombination, one or more functional site sequences that regulate the gene. This procedure may be used, for example, to reduce the expression of a gene that is inappropriately being overexpressed in a patient and causing disease. Such procedures may be performed *ex vivo* in affected cells, *e.g.* bone marrow

cells, which may be implanted back into a patient. In addition, such methods may be used to create animal models of human disease, *e.g.,* diseases caused by reduced or altered expression of a gene.

In other embodiments, the invention provides knockdown methods to reduce expression of a gene, which involve introducing a knockdown reagent targeted to a sequence of the invention. Such methods are particularly useful when the functional site sequence is located within a transcribed region of the gene.

In a variety of embodiments, the invention provides methods of regulating or providing exogenous gene expression using polynucleotide sequences of the invention. Such exogenous genes are typically expressed from cDNAs, and may correspond to genes and polypeptides normally found within the cell or organism or genes or polypeptides foreign to the particular cell or organism. In addition, in certain embodiments, such exogenous genes may encode polypeptides, while in other embodiments, they may not encode polypeptides, but, rather, they may encode functional nucleic acids, such as, antisense, ribozymes or shRNA molecules, for example. These methods include, but are not limited to, methods involving transient expression vectors, transgenic methods, and methods of gene therapy. Such methods may be employed to direct expression of a gene or nucleic acid in any cell or at any time, depending on the functional site used. Thus, functional site sequences may be used to drive expression of an exogenous gene constitutively or inducibly, globally or in a particular cell or tissue type, for example, by placing the gene under the control of a functional site active in the desired cell or tissue type, at a particular developmental or cell-cycle stage. In addition, such methods may be employed to increase or reduce expression of a gene or nucleic acid. Furthermore, genes or nucleic acids may be regulated according to the invention *in vitro* or *in vivo*, and the expression, knockout, or knockdown constructs may be either transiently present within a cell or stably integrated into a cellular genome.

Accordingly, in certain embodiments, functional site sequences of the invention are used for experimental or therapeutic control of transcriptional programming. Polynucleotides comprising functional site sequences may be used to design molecules that can interfere with the formation of a functional site in the nuclei and so control transcriptional regulation. The inhibition of the formation of a specific hypersensitive site may cause an expected alteration in the transcriptional program or induction of a different pattern of active genetic elements. Since functional sites may be associated with increased or decreased transcription, such methods may be employed to decrease or increase expression of a gene, particularly a gene regulated by the targeted functional site. As an experimental tool, functional sites can, thus, be used to perform functional gene knock out experiments or otherwise examine the redundancy of the regulatory network in the nucleus, *in vitro* or in *vivo*. Molecules that may be used according to the methods of the invention to interfere with the formation of a functional site include, *e.g.*, antisense and other knockdown reagents, competitive oligonucleotides, inhibitory antibodies directed to functional site binding proteins and dominant negative polypeptides corresponding to regions or mutants of functional site binding proteins.

The invention further provides methods of stimulating methylation at functional sites by introducing a complementary and methylated polynucleotide corresponding to the targeted functional site. At functional sites, a strong correlation exists between demethylation of certain sites (of the cytosine in CpG dinucleotides) and increased transcriptional activity. These key CpG dinucleotides can be re-methylated by introduction of a complementary polynucleotide containing a 5-methylcytosine at the crucial position; the resultant hemi-methylated site will be a substrate for the maintenance methylase activity present in eukaryotic cells. The introduction of a methylated CpG dinucleotide into the site would be expected to change its transcriptional influence and may be accomplished, for example, by homologous recombination methods.

The invention provides a variety of methods of identifying functional sites associated with or involved in the regulation of a specific gene. In one embodiment, a functional site associated with a gene of interest is identified by identifying a functional site with a genomic location near a gene of interest. The invention provides the genomic location of the functional sites described here, thereby proving the means to identify functional sites located within any gene with a known genomic location. The ability of the identified functional site to regulate expression of the gene may be confirmed as known in the art, *e.g.* by knocking out or knocking down the functional site in a cell or animal and determining the effect on expression of the gene by comparing expression in a knockout or knockdown cell to a wild type or vector control cell. Transgenic studies have been used to demonstrate that functional site sequences required to recapitulate developmentally- and tissue-specific correct expression of genes are contained within discrete fragments of genomic DNA. The best characterized system is that of the 248 kb YAC comprising a genomic fragment containing all the human beta-globin genes and the sequences necessary for their correct expression (Peterson *et al.*, Proc. Natl. Acad. Sci. USA 90: 7593). This study and others like it (reviewed in Li *et al.*, Blood 100: 3077) demonstrate that the proximity of functional site sequences to a gene sequence implicates it in the regulation of the gene. Using this simple parameter, the skilled artisan could readily identify functional site sequences likely to be necessary and sufficient to effect regulatory control on any given gene. Functional sites associated with a specific cell or tissue may be used alone or in combination to direct expression of a therapeutic gene to a desired cell or tissue type. In addition, appropriate expression of a therapeutic gene, *e.g.*, in replacement therapy, may be accomplished using a gene therapy construct wherein expression is regulated by one or more functional sites with genomic locations near the corresponding endogenous gene.

As briefly described above, functional site sequences of the invention may also be used for transgenic expression of genes. In fact, the identification of

functional sites necessary for either proper expression of a particular gene or with generic properties that will contribute to high-level expression of genes in general is very important in this new field. For example, some of the sequences form chromatin insulators and boundary elements and are useful for transgenic studies to drive proper and proportional expression of transgene. Approaches such as transgenic therapy often require engineering of a construct that is capable of being delivered to the nuclei and driving appropriate expression of the introduced transgene. In this context, appropriate can mean directing a high level of expression of the transgene only in the tissues desired. This may be achieved by including the correct set of regulatory elements, *i.e.* functional sites in the transgenic construct, and including insulator elements to buttress the transgene against the generally repressive effect of the bulk chromatin into which it has inserted (Emery *et al.*, Blood 100: 2012). Insulators may also serve to restrict the influence of the construct's regulatory sequences to the transgene, so as to not influence the expression of any genes neighboring the insertion event.

The invention further provides methods of identifying a gene associated with a particular phenotype, such as a specific differentiated or cell cycle state, or example. The invention provides functional site sequences specifically associated with different cells. Genes regulated by such functional sites may be identified based upon their physical proximity to one or more of such functional sites by comparing the genomic location of the functional site to a genome map and identifying genes close to the functional site. In certain embodiments, genes regulated by functional sites are located within 100 bp, 500 bp, 1 kb, 2 kb, 5 kb, 10 kb, 100 kb, or greater than 100 kb of the functional site.

2. Diagnostic and Therapeutic Methods

In certain embodiments, functional sites serve as therapeutic targets. Since functional sites regulate expression of genes, they may be targeted to alter gene expression in a therapeutic manner. In one embodiment, the invention

provides a method of reducing expression of a deleterious or disease-associated gene by targeting or altering the activity of a functional site that regulates expression of the gene. A variety of sequence-specific agents, including the knockout and knockdown reagents described *supra*, may be used to alter expression of a gene regulated by the targeted functional site. An inevitable extension of this (or an analogous) chemistry will be an ability to target particular DNA regulatory regions in living tissues. The sequences may be used for designing complementary polynucleotides to interfere with the formation of functional sites. Some functional sites stimulate disease either through their formation or from their absence. The sequence information associated with such sites can be used to design molecules, such as polynucleotides or synthetic chemicals, to block their formation in nuclei. Additionally, sequence-specific DNA 'nano-binders' such as polyamides are an established research area and are currently being developed as pharmaceuticals. Furthermore, since functional sites typically function by binding one or more polypeptides, functional site activity may be altered by expressing or otherwise introducing a dominant-negative polypeptide corresponding to a polypeptide that binds a functional site, directly or indirectly. In one embodiment, such dominant-negative polypeptides will retain their DNA binding another activity necessary for functional site activity, such as the ability to bind a coactivator, for example. Inhibition of the formation of some functional sites can have a similar effect on the development of the disease phenotype.

Clinical studies of a certain disease state contain populations of individual genomic samples from patients verified to have the disease or not. Functional sites may be identified in the genomic DNA isolated from patients with the disease or control samples without the disease, and sites associated with the disease identified. Such information may be used diagnostically to determine whether an individual has a disease by determining whether the individual has functional sites specifically associated with the disease. In one embodiment, the invention provides an array based screening system for detecting individual

functional sites or patterns of these that are associated with the disease state and as such be useful in generating diagnostics.

Accordingly, the invention provides method of using the functional site sequences of the invention to detect or diagnose the presence of a disease or disorder in a patient. For example, by comparing the hypersensitivity sites present in a diseased cell or tissue to those present in a normal or non-diseased cell or tissue, sites correlating (*e.g.* present or absent) to the presence of any disease may be identified. Cells from a patient suspected of having a disease may then be examined or profiled to identify the presence or absence of one or more hypersensitivity sites associated with a disease or disorder. The presence or absence of a hypersensitivity site, as associated with a specific disease, indicates that the patient has the disease. In certain embodiments, hypersensitivity sites from a patient are determined and compared to databases or computer readable medium comprising sets of hypersensitivity sites associated with a disease to determine if the patient has a disease. The patient's hypersensitivity site profile may be compared to profiles established for one or a plurality of diseases. Therefore, the method may be used to detect or diagnose disease in the absence of clinical symptoms or any other indication of the nature of the disease.

Many of the sequences are associated with hypersensitive sites involved in disease. Labeled (by fluorescent dyes for example) polynucleotide probes (such as complementary PNA) or synthetic molecules designed to recognize stretches of the highly accessible regions of the DNaseI hypersensitive sites can be used to detect their formation in intact nuclei. The detection of those hypersensitive sites associated specifically with disease states either by studying nuclei which have been isolated or are still intact could be used to detect, evaluate and monitor cells with a regulatory environment associated with diseases.

Sequences and sites identified using one or more of the sequences listed in Table 1 (*e.g.*, genes regulated by the sequences set forth in Table 1) can be used in the diagnosis and treatment of diseases and disorders. A diagnostic

may comprise a sequence (which may be DNA, RNA or PNA) coupled to a solid support for the detection of a complementary sequence (which may be DNA or RNA). Levels of expression (or trends of expression over time) of the complementary sequence can be determined from a biological sample obtained from a patient (*e.g.*, DNA, hnRNA, mRNA, rRNA, miRNA, ncRNA, stRNA, RNAi) as a disease indication. Expression levels significantly lower or higher in the test sample as compared to expression levels in a normal control sample may indicate the presence of a disease or disorder. In certain embodiments, a reference value is determined based on the expression levels in one or more normal controls, and the presence of a disease or disorder is determined by comparing expression levels in the test sample to the reference value. In certain embodiments, a two-fold difference in expression is considered significant, while in other embodiments, a three-fold, four-fold, or five-fold difference is considered significant. The skilled artisan would readily appreciate that normal levels of mRNA and polypeptide expression may fluctuate or vary between different normal controls, and will take such variation into account when determining an appropriate reference value and a significant level of variation from a normal value.

In certain embodiments, nucleic acids of the invention may be used to identify a gene associated with a disease or disorder. For example, functional sites present in a diseased cell or tissue may be identified and compared to those present in a normal cell or tissue. Functional sites either present or lacking, or present to a differing degree, in the diseased cell or tissue are considered to be associated with the disease or disorder. The genomic location of one or more differing functional sites may be determined and a gene located near to the site or known to be regulated by the site may be identified as involved in or associated with the disease. One of several means of confirming the relationship of the identified gene and the disease is to measure the levels of mRNA or polypeptide expressed from the gene and compare it to the levels observed in normal cells. Any different would confirm that the gene is associated with the disease or

disorder. Methods of measuring mRNA and polypeptide levels are widely known and available in the art and include, for example, RT-PCR and western blotting, for example.

In another embodiment, the presence of a disease or disorder may be determined by sequencing one or more nucleic acid sequences of the invention and identifying a mutation or sequence aberration in a functional site of a patient as compared to a normal control. In another embodiment, the presence of a disease or disorder may be determined based on differences in cleavage by a nuclease, chemical, or other agent used to detect functional sites.

The invention further provides methods of screening candidate drugs and other potentially therapeutic compounds, as well as assessing the efficacy of therapies. The ability to screen an entire set of functional sites allows the development of screening protocols that not only define the reaction to application of a drug or treatment on a defined set of sites or parameters but to a global set. In this manner, drugs and chemicals that induce non-specific effects are detected.

The sequences may be used for toxicological profiling of potential drugs. Characterizing the molecular consequences of applying or titrating a drug into cell populations, tissues or test organisms is very useful to define the pathways and side effects of a drug. Comparison of the patterns from hybridization experiments using the isolated hypersensitive sites probed with the probes derived from the test populations can confirm the mechanism of the drug. Testing the response of patients to a regime of drugs also allows identification of patients who may be more or less suitable for that particular treatment, as some patients may show little induction of the target active genetic element or an unexpected activity in other sets of hypersensitive sites. Thus, in one embodiment, the invention provides a method of qualifying a patient for a clinical trial or for treatment with a drug or therapy that involves determining the hypersensitivity profile of the patient and comparing it to the functional site profiles of patients known to respond positively or negatively to a particular drug or

therapy. Alternatively, the status of one or more individual functional sites may be used for such purposes.

In a related embodiment, the invention provides a method of correlating clinical data with functional sites to predict the outcome of a disease or treatment protocol. Functional site profiles may be established for patients and correlated with disease progression or outcome, alone or after treatment with any therapy or protocol. The functional site profile may then be determined for a patient and used to predict disease outcome or the success of a given treatment protocol and will assist in determining the appropriate therapy.

The sequences may also be used for discovery of novel lead compounds. Drug discovery can be advanced by understanding the biology of the target disease system and in particular identification of functional sites involved in disease progression. High throughput screening, using labeled probes able to detect the formation of hypersensitive sites in nuclei, can be used to identify compounds that affect the formation of specific functional sites.

Functional site profiles may be compared between cells treated with a drug and untreated or control cells to identify drugs and drug candidates. Furthermore, where a specific functional site has been associated with a disease or disorder, probes specific for such site may be used to determine the status of the site before and after drug treatment to identify a drug that alters the status of the functional site and would, therefore, be useful in therapy of the associated disease or disorder. In certain embodiments, treatment with the drug restores the status of the functional site to that observed in a control sample.

Functional sites and profiles thereof may also be used according to the invention for toxicology profiling of drugs. Determination of the effect of a drug upon hypersensitivity sites may be predictive of drug toxicology, for example, based upon the effects of known toxic agents or drugs upon one or more functional sites. In another embodiment, drug toxicity may be correlated with specific patients based upon the presence or absence of one or more functional sites in

patients wherein a drug has been toxic. The ability to predict drug toxicity, particularly where only a relatively small number of potential patients are susceptible, will allow physicians to selectively avoid treating patients potentially subject to drug toxicity.

The invention further provides drugs identified according to any of the methods of the invention. For example, the invention provides drugs, including small molecules, for example, identified as effecting or altering the accessibility of on or more functional sites. Accordingly, the invention provides a drug produced by the process of screening one or more compounds for their ability to alter one or more functional sites, identifying a compound that alters one or more functional sites, and producing said compound. Alterations in functional sites may be detected based upon changes in their cleavage or accessibility by nucleases or other agents that cleave DNA, for example, and may involve, *e.g.,* either an increase or a decrease in hypersensitivity.

The invention further provides methods of manufacturing a drug of the invention. Such methods comprise identifying a drug that effects or alters one or more functional sites of the invention and producing the identified drug.

3. Identification and Characterization of Cells

In certain embodiments of the invention, functional sites may be used to identify specific cells. Functional sites need not solely be used as a marker of diseased cells. Rather, they may be used as a marker of any cell. For example, functional sites associated with a specific cell may be identified and used to establish a unique identification pattern analogous to a genomic fingerprint associated with the cell. In certain embodiments, the cell is a specific cell or tissue type, at a particular developmental stage, associated with a particular disease or disorder, or treated or exposed to an environmental stimuli or drug, for example. An example of another application would be to define sites as a function of developmental progress. Such markers would be useful in definitively and

quantifiably determining the stage of cells within a population and may be harnessed to effect cell sorting.

The sequences may be used as *in vivo* markers for classification and sorting cell populations. The formation or presence of functional sites crucial for induction of certain genes will define the position at checkpoints of each cell in terms of its developmental progress and tissue-specificity. Using labeled probes directed towards functional sites, which remain inaccessible when the site is not formed, will allow detection of such 'markers' in intact nuclei. By use of, for example, two fluorescently labeled probes which give a strong FRET signal when bound to the same region of a formed functional site, it will be possible to fractionate (using FACS) a population of cells from complex mixtures according to any criteria, such as, for example, their exact developmental stage or tissue-description.

The sequences may be used for functional tissue typing. The ability to detect formation of functional sites in nuclei allows construction of a regulatory profile for individual tissues or mixtures of tissue, either separated from primary tissue or from monocultures. A thorough understanding of how these profiles change due to a stimulus, such as drug treatment, allows the isolation of cells from a previously homogenous population, which are highly potentiated. An example is the sorting of totipotent stem cells from a larger population or stem cells that have successfully been pushed down a particular developmental pathway.

4. Identification of Regulatory Proteins

The invention further provides a series of technologies to identify the proteins that interact with the identified functional site sequences. The fact that functional sites have been shown to have 'core' sequences, which are necessary and sufficient for their formation, if not necessarily their biological activity, suggests that there are protein-protein interactions crucial for functional site formation (Stamatoyannopoulos *et al*, EMBO J. 14: 106). The identification of such critical

interfaces in the formation of regulatory complexes represents the discovery of a new generation of therapeutic targets, *e.g.,* to competitively complex proteins *in vitro* or *in vivo.* Such activity has been demonstrated and is in late-stage clinical trials for DNA sequences that bind certain transcription factors – so-called 'decoy oligonucleotides' directed against a transcription factor, E2F.

The sequences may be used for discovery and analysis of DNA-protein interactions. The identities of proteins participating in the interactions and their functions may be determined. For example, key proteins involved in transcriptional regulation maybe identified. Polynucleotides capable of binding to functional sites may be identified by any of a variety of means known and available in the art, including the use of binding columns comprising the functional site polynucleotides or fragments thereof and yeast-based screening assays, for example. Polynucleotides comprising functional sites can be labeled and used as substrates in electro-mobility shift assays (EMSA) to identify, which proteins from a range of nuclear extracts bind to the sequence. Addition of antibodies raised against candidate nuclear proteins can be used to cause a further 'super' shift allowing identification of the individual protein components within the nucleo-protein complex. Once the nature of the proteins are known the presence of post-transcriptional modifications can be determined, by use of antibodies raised against specific modifications or by mapping the mass of fragments by mass spectrophotometry, for example.

The sequences may be used as templates for *in vitro* or *in vivo* footprinting and to identify the binding positions of DNA-binding proteins. Footprinting of the cloned sequences maybe carried out with a variety of cutting agents, such as DNaseI or free radicals, for example, which reveals patterns of binding of proteins either *in vitro* to a panel of nuclear extracts or purified components or *in vivo* in different tissues. The binding of a particular protein is specific to its cognate site, many of which are known and hence can be used to infer the proteins bound to the functional site. The region of the functional site that

the protein covers can indicate the overall structure, and therefore function, of the functional site.

The sequences can identify proteins bound to and associated with functional sites. The identification of all of the components of the hypersensitive sites *in vivo* is made possible by hybridizing nucleic acid having sequence(s) of cloned functional sites to exposed regions of fractionated chromatin. For example, cross-linked sonicated chromatin can be treated with exonucleases to expose single-stranded DNA regions that can form targets for biotinylated fragments from the cloned functional sites. Such captured complexes can be analyzed for protein content and levels of epigenetic modifications. In this example, both protein-DNA and protein-protein interactions can be determined. The available techniques for carrying out these studies potentiate the discovery of interactions between functional sites, as proposed by looping models, wherein transcriptional enhancers interact with their cognate promoters via complex protein-protein interactions. The existence of such complexes may be a general effect or may be restricted to a number of super-regulatory elements or LCRs (Locus Control Regions).

Yet another use of the sequences is for the study of post-transcriptional modifications within the genome operating system. The CHIP protocol (chromatin immuno-precipitation) has been used to enrich for sequences, often from formaldehyde cross-linked nuclei, bound by nuclear proteins or by proteins carrying post-transcriptional modifications, such as the acetylation pattern of histones. This pool of fragments can be used to hybridize to the isolated functional site sequences to determine, for example, which functional sites are bound by which nuclear protein. In the case of post-transcriptional modification, the changes in these epigenetic markers can be followed as a function of tissue-type and development.

The sequences may be used to probe the role of differential methylation within functional sites. Analysis of the sequences of the cloned functional sites can reveal the presence of CpG-dinucleotides. Some of these

dinucleotides can be differentially methylated at cytosines, and such methylation sometimes causes transcriptional inactivity of an associated gene. Genomic sequencing can be used to compare the methylation status of such potential epigenetic modifiers across a panel of nuclei in an attempt to identify those that may have key regulatory functions.

The sequences may be used as markers for studying the role of nuclear localization in transcriptional induction. It is possible to follow the nuclear localization of specific sequences using fluorescently labeled probes and confocal microscopy. The existence of sub-compartments within the nucleus and the recruitment of functional site sequences and genes to them potentially plays a major role in understanding transcriptional regulation in eukaryotic nuclei. Most preferably, a panel of labeled probes is generated against sets of functional sites. The distribution of the sites is monitored throughout the nuclei and compared with different systems or under different conditions

In another embodiment, sequences are used for raising antibodies against components of isolated functional site complexes. Successful isolation of the intact nucleoprotein complexes, *e.g.*, by hybridization with biotinylated sequences derived from the cloned functional sites allows the generation of monoclonal and polyclonal antibodies against both the proteins bound in the complex and the tertiary structure of it. Such antibodies are useful in a range of applications such as CHIP, wherein antibodies raised against the nucleoprotein complex as a whole have higher specificity. The antibodies also may be used in studies that disrupt the function of the functional site by interfering with molecule(s) that interact with the functional site in its natural context.

Nucleoproteomic analysis will also detect novel transcription factors or co-activators associated with pathogenic functional sites. In addition, mass spectrophotomeric analysis may be utilized to detect *a priori in vivo* post-translational modifications of such proteins. Previous studies have shown that

such modifications may be crucial in proper regulation of gene systems (Song *et al.*, J. Biol. Chem. 277: 7029).

5. Identification of Genomic Modifications

The invention further provides methods of identifying mutations, *e.g.* allelic variants, associated with disease. The complete set of functional sites, though representing a small percentage of the genome, will contain much of the regulatory-significant genetic variation (such as SNPs). In one embodiment, the invention provides array systems that allow the simultaneous detection of the presence of sequences corresponding to one or more functional sites in a patients genomic DNA and the detection of genetic variation or mutation, which involve placing short versions of the targets differing in the sequence of a single base between them on the array (referred to as 'array based sequencing').

A skilled artisan using the sequence information provided in Table 1 may conveniently identify genetic anomalies associated with one or more human diseases (*e.g.*, cancers, immune disorders, neurological disorders, cardiac disorders, or genetic disorders generally). For example, by matching known genetic changes associated with malignant transformation with the precise sequence or position information of a functional site, it is possible to identify genetic anomalies in functional sites associated with a specific cancer. In another embodiment, functional sites associated with a specific disease are identified by comparing functional sites and/or their genome location identified in a disease sample as compared to those identified in a normal control sample. Hypersensitivity sites specifically associated with a disease may be used individually or as a set to detect or diagnose the disease or to identify regulated genes involved in disease onset or progression, for example.

In one embodiment, functional site sequences are used to map disease-causing SNPs (single nucleotide polymorphisms). Such single nucleotide polymorphisms, which cause changes to the expression pattern of nuclei, are more

frequent within active genetic elements. *A priori,* the database of known functional sites may be screened to capture a population of phenotypically active SNPs.

The information presented in or obtained from Table 1 herein provides a template for deriving new, previously unknown functional site sequences for known genes. In one embodiment, an active genetic sequence for a protein-coding gene is discovered or further characterized by comparing the positional information shown in Table 1 with the known location of the gene. For example, sequence information obtained from Table 1 can be used to design primers for polymerase chain reactions (PCR). A functional site sequence that is close (preferably within 10,000 base pair, more preferably within 3,000, 1,500 or 500 base pairs) to a protein encoding gene, single nucleotide polymorphism (SNP), or other site of interest, may be selected by a computer. Sequences for primer recognition can be selected and PCR reactions performed to identify and/or quantitate SNPs, changes in chromatin structure, or genome-specific mutations or individual-specific mutations. In an embodiment, a protein-encoding gene already has a known regulator that may be similar in location to a functional site sequence. Information in Tables 1 is used to discover a further attribute of the known active genetic sequence, such as the location of a functional site that may be at the edge or outside the known regulator. In the latter case, this embodiment of the invention allows the discovery of a new section or border of a previously considered gene regulator. In yet another embodiment, multiple functional site sequences that affect the same gene or set of genes are discovered by virtue of their clustering in genome space.

In one embodiment, a pre-existing set of genetic changes associated with a disease is compared with information from Table 1 to determine which of the changes linked to disease involve functional sites or regulatory DNA. This information provides great value for drug discovery and for new modalities for treating disease.

In yet another embodiment, allelic variants of regulatory DNA sequences corresponding to Table 1 are correlated with genetic diseases. Allelic variants may be identified, for example, by comparing the accessibility at a specific site from samples derived from different individuals or cells, including individuals or cells with a disease, for example. Such comparisons may be based upon alterations in accessibility or digestion at the site, ability to bind regulatory molecules, the presence or absence of chromatin modification such as methylation or acetylation, for example, chromatin or, alternatively, the sequence of a site.

A set of functional sites can be considered as a panel of markers across the genome and as such have a wider distribution than that of traditional expression arrays. Performing hybridizations to arrays comprising functional site sequences using probes prepared simply from sonicated genomic DNAs has the potential to detect different patterns of deletions, expansions and duplications between different genomic DNAs, in other words to construct karyotypes. Furthermore, high copy number of DNA viruses may be detected in this manner.

D. Databases and Computer Programs

The invention further provides materials comprising information, which are useful for a variety of purposes, including the identification of a cell or tissue type, the identification of regulatory sequences, the identification of disease-associated sequences, and regulating gene expression. In certain embodiments, such materials include information libraries and databases, which may be printed or stored on computer readable media, for example.

In many embodiments, the invention provides information libraries related to functional sites of the invention, which may comprise the nucleic acid sequence of functional sites, the genomic location of functional sites, diseases associated with functional sites, or genes associated with or regulated by functional sites, or any combination of such information. In addition, information libraries of the invention may include sets or subsets of functional site sequences

and related information, including, *e.g.,* functional sites associated with a specific cell, tissue, disease, or gene. Information libraries may also comprise functional site variant sequences, such as SNPs, including those associated with a disease or other characteristic. Such information libraries, particularly on computer readable medium or other searchable formats, provide valuable databases that may be searched, *e.g.,* to identify functional sites associated with a genomic locus or gene, to identify one or more functional sites that regulate a gene, to identify mutant or variant functional sites associated with a disease or disorder, and to identify a gene associated with a mutated or variant functional site.

Sets of sequences and/or positional locations may be prepared by computer analysis of information provided herein and have great intrinsic value for a variety of uses such as regulatory unit discovery, diagnostics and therapeutics. Particularly contemplated are sets of functional site sequences and/or genomic positions that correspond to regions of the genome, such as particular chromosomes, hypervariable regions that experience high levels of DNA breakage, and the like. After computer formation, such sets of data, presented in computer readable form or directly readable by a person, are valuable items of commerce and may be sold directly.

Thus, in one embodiment, the invention provides computer-readable medium comprising or consisting of a plurality of nucleic acid sequences of functional sites of the invention, including, for example, functional sites identified in human fetal brain genomic DNA or K562 cell genomic DNA. In one embodiment, the computer readable medium comprises or consists of a plurality of sequences set forth in Table 1. In certain embodiments, the computer-readable medium further comprises the genomic location of and/or the identification of genes regulated by the nucleic acid sequences. In another embodiment, the computer-readable medium comprises a plurality of functional site sequences associated with a disease, disorder, or clinical outcome, a specific cell type, a specific transcription factor, treatment with a drug or drug candidate, or one or more

chromosomes or genes, for example. The computer-readable medium may also comprises the genomic location of and/or identification of any genes regulated by the nucleic acid sequences.

In the context of arrays of nucleic acid sequences of the invention, a computer-readable medium may include any of the information described *infra*, and, in addition, may further comprise the location of each arrayed nucleic sequence on the array. The computer-readable medium may, therefore, provide the sequence of a nucleic acid located at a specific location on an array. The computer-readable medium may further provide any other known information about the sequence at a specific location on an array, including, for example, the genomic location of the sequence, any genes associated with the sequence, transcription factors associated with the sequence, and any diseases or disorders associated with the sequence, etc.

The invention further provides computer programs and software useful in the analysis and compilation of the functional site data. In many embodiments, a computer program is used that inputs at least part of the functional site sequence and other information provided herein, such as at least 10, 100, 1,000, 10,000 or more sequences and genome locations and then selects out a smaller set therefrom.

As one example, a software program can direct a computer to find allelic forms of a functional site by searching public data bases for sequences of functional sites and variants thereof, based upon the genomic location information provided by the invention, that may be input into the computer. An allelic form of a functional site may be identified, for example, by inputting the genomic location and/or sequence of the functional site, searching a public or private database comprising genomic sequence information, and identifying an allelic form based upon it having the same genomic location and a variant sequence as the inputted functional site.

In one embodiment, a computer under direction of a program inspects the genome location contents of a database provided by the invention and chooses a functional site near a desired gene, thereby identifying a functional site associated with the gene, determining a previously unknown regulatory unit, identifying a new function for an under appreciated functionally functional site, or providing greater clarity as to the borders of a known regulatory unit. According to this embodiment, preferably the computer looks for functional sites within 100,000 base pairs of a selected gene, and more preferably within 20,000; 10,000; 3,000; 2,000; 1,000; 500; 200; 100; or even 50 base pairs of the selected gene.

The DNA sequences and their location information provided herein may be used for further discoveries through data mining, using a portion, or all of the listed information. For example, the genomic location information reveals clustering of functional sites in the genome, as can be readily apprehended by a computer directed by a program to group functional sites of the invention that physically locate close to each other in the genome. Generally speaking, such grouped sites, termed "clusters" regulate coordinately one or more genes.

In a desirable embodiment, a software program instructs a computer to load multiple genome locations of functional sites and then compare how far apart each genome location is from the others. The program instructs the computer a set maximum genome distance for comparison and to decide if two sites are less apart than that distance. If they are, then this fact is noted in memory, such that the two are labeled or grouped into the same cluster. Most conveniently, a cluster will be made by storing identifiers for the functional sites at the same or adjacent areas of memory. Cluster groups may be stored on long-term media (*e.g.*, hard drive, CD ROM) and/or displayed to the computer operator. In an embodiment two functional sites are deemed within the same cluster if their genome locations are within 1,000,000 bases of each other. In another embodiment, functional sites are deemed part of a cluster if their genome locations

are within 300,000 bases, 100,000 bases, 30,000 bases, 10,000 bases, 3,000 bases 1,000 bases or even 250 bases of each other.

For purposes of brevity, a separate listing of each possible subset of sequences contemplated is not presented, and space limitations are overcome by the convenient use of computers to group the data as described herein. Thus, specific clusters of functional sites, found on each chromosome, and selected by closeness based on proximity are intended embodiments and can be easily printed in tabular form as desired with the aid of a computer.

One embodiment of the invention provides a computer program that determines a cluster by reviewing the genome positions of multiple functional sites (at least 100, 1,000, 10,000 or more) and placing sites having near positions to each other within one or more of the above specified ranges, into a common group. This embodiment of the invention is made possible by the fact that the information set forth herein was obtained under real conditions wherein multiple coordinating functional sites were actively controlling gene expression. That is, the sites listed in Table 1 are not a random assortment of functional sites in the genome and do not necessarily represent all possible sites, but represent functional sites that were simultaneously active in a functioning cell system. Among other things, this property distinguishes the information provided by the invention from other data sets obtained by others using purely computer analysis of the sequenced genome.

In another embodiment, a software program instructs a computer to compare known sets of genetic changes associated with a disease with functional site sequences of the invention. The computer inputs at least one set of genetic information, inputs at least some sequence information and/or genome positions from the information provided, and compares identities using a known algorithm or procedure. After comparing the two sets, the computer selects a match set to be output or used for further analysis, indicating one or more sequences associated

more definitively as functional sites or regulatory regions of more defined sequence and size.

Sets of members selected from the functional sites of the invention may be prepared and used as articles of commerce, research tools, diagnostic aids, drug discovery aids and the like, based on a desirable grouping category such as those based on genetic changes in malignancy and genetic changes associated with specific disease. In a very desirable embodiment, a known genetic abnormality is used to find linked functional sites that cooperatively influence gene expression or an overall biological process mediated by multiple genes. This is carried out by examining for unknown cluster partners. In this embodiment, functional sites associated with a known DNA problem such as a disease or allelic form of a gene associated with a definable trait based on, for example, an improper transposition, deletion, or other mutation, are placed into a set, and combined with further members that are found to cluster with the known genetic errors.

Software programs are contemplated for discovery and use of these sequences and positional information. In each case, a set of functional site sequences, and/or positions in the human genome are loaded into a computer and stored, in volatile memory, short-term erasable memory and/or long term non-erasable media. A program is loaded into the computer that parses through the set of sequences and/or genomic positions. For each parse, the computer makes a decision having biological or biochemical relevance. For example, one type of decision is to determine whether a parsed sequence is similar to (homogenous to) a known functional site sequence such as a known promoter or so-called "enhancer" sequence. The computer may look for strict equivalency in sequence of course but in many embodiments the computer will examine for a minimum percent homology or other correspondence as is known in this art. By way of example, if a segment of a sequence of about 15, 20, 30, 50, 75, 100, or 200 bases of those shown in Table 1 is at least 70%, 75%, 80%, 85%, 90%, 95%, or at

least 97% identical to a reference known functional site sequence, the computer will store the correspondence information or match in memory for use later by a program or for display to the computer operator. In most embodiments, the computer will store selections in memory and later transmit a set of selections by electronic transfer to a permanent medium such as an optical or magnetic disk or by electronic transmission.

The invention further includes the information gathered by the previously described computer programs, including, for example, the genomic position of identified sequences. Such information includes data and data sets in both printed and computer-readable format. Thus, the invention provides computer readable medium comprising a plurality of nucleic acid sequences identified as functional sites. The data may be further defined or classified to include functional sites associated with specific cells or tissues, diseases, chromosomes, transcription factors, or chromatin structure or modification, for example. Indeed, the skilled artisan would readily appreciate that data or computer-readable medium containing functional site sequences from any cell of interest, including for example, a cell treated with a drug or drug candidate, is contemplated by the invention and of value in identifying genes and gene regulation associated with the cell.

E. PHARMACOGENOMICS

A further use for the functional sites of the invention is for evaluating differential efficacy of or tolerance to a treatment in a subset of patients who differ in genotype with respect to one of more functional sites listed in Table 1. The patient can be human or non-human, for example a rat, mouse, dog, cat, or non-human primate.

The patient can differ in genotype of a functional site relative to the sequence listed in Table 1 in either a single allele or in both alleles of the functional site.

In certain embodiments, the stratification is based on one or more polymorphisms in a single functional site. In other embodiments, the stratification is based on one or more polymorphisms in two or more functional sites. The two or more functional sites can be selected on the basis of their association with a single gene involved in a disease state, or their association with two or more genes involved in a disease state.

Thus, the patient stratification methods of the present invention allow the identification of selecting a suitable dosage level and/or frequency of administration, and/or mode of administration of a compound for a patient's regulotype with respect to one or more of the functional sites listed in Table 1. The method of administration can be selected to provide better, preferably maximum therapeutic benefit.

The functional site alleles or polymorphisms that can serve the basis for stratification of patients can be an insertion, deletion, substitution, or a combination of two or more of the foregoing relative to the functional site reference sequence provided in Table 1 herein.

In preferred embodiments, the correlation of patient responses to therapy according to patient's functional site genotype is carried out in a clinical trial. Alternatively, the information from clinical trials previously conducted can be reassessed if the patients' functional site genotypes can be evaluated. Thus, for example, patients may be stratified by genotype and the response rates in the different groups compared, or patients may be segregated by response and the genotype frequencies in the different responder or nonresponder groups measured. One or more functional site polymorphisms may be studied.

Stratification of patients in clinical trials according to the methods of the invention can result in the identification of a subset of patients with enhanced or diminished response or tolerance to a treatment method or a method of administration of a treatment where the treatment is for a disease or condition in the patient.

The method involves correlating one or more polymorphisms in one or more functional sites in Table 1 in a plurality of patients with response to a treatment or a method of administration of a treatment for a disease with which the functional sites are associated. The correlation may be performed by determining the one or more polymorphisms in the plurality of patients and correlating the presence or absence of each of the polymorphisms (alone or in various combinations) with the patient's response to treatment. A positive correlation between the presence of one or more polymorphisms and an enhanced response to treatment is indicative that the treatment is particularly effective in the group of patients having those functional site polymorphisms. A positive correlation of the presence of the one or more functional site polymorphisms with a diminished response to the treatment is indicative that the treatment will be less effective in the group of patients having those variances. Such information is useful, for example, for selecting or de-selecting patients for a particular treatment or method of administration of a treatment, or for demonstrating that a group of patients exists for which the treatment or method of treatment would be particularly beneficial or contra-indicated. Such demonstration can be beneficial, for example, for obtaining government regulatory approval for a new drug or a new use of a drug.

The methods of the present invention encompass identifying a first patient or set of patients suffering from a disease or condition whose response to a treatment differs from the response (to the same treatment) of a second patient or set of patients suffering from the same disease or condition, and then determining whether the occurrence or frequency of occurrence of at least one functional site polymorphism that differs between the first patient or set of patients and the second patient or set of patients. A correlation between the presence or absence of functional site polymorphism and the response of the patient or patients to the treatment indicates that the polymorphism correlates with patient response. In general, the method will involve identifying at least one polymorphism in at least one functional site that correlates with a patient's response to a treatment.

The methods of the invention can utilize a variety of different informative comparisons to identify correlations. For example a plurality of pairwise comparisons of treatment response and the presence or absence of at least one functional site polymorphisms can be performed for a plurality of patients. Likewise, the methods can involve comparing the response of at least one patient homozygous for at least one functional site polymorphism with at least one patient homozygous for the alternative form of that polymorphism. The method can also involve comparing the response of at least one patient heterozygous for at least one functional site polymorphism with the response of at least one patient homozygous for the at least one functional site polymorphism. Preferably the heterozygous patient response is compared to both alternative homozygous forms, or the response of heterozygous patients is grouped with the response of one class of homozygous patients and said group is compared to the response of the alternative homozygous group.

In an another related aspect, the invention provides a method for identifying a patient for participation in a clinical trial of a therapy for the treatment of a disease listed in Table 1. The method involves determining the genotype or haplotype of a patient's one or more functional sites associated with the disease. Patients with eligible genotypes are then assigned to a treatment or placebo group, preferably by a blinded randomization procedure. In preferred embodiments, the selected patients have two copies of the reference functional site sequence listed in Table 1, one copy of the reference functional site sequence listed in Table 1 and one copy of a different allele, or two copies of a functional site allele that differ in sequence relative to the functional site reference allele listed in Table 1. These two copies can be of the same or of different alleles. Regardless of the specific tools used to group alleles, the trial would preferably test for a statistically significant difference in response to a treatment between two groups of patients each defined by their functional site genotype. Said response may be a desired or an undesired response. In a preferred embodiment, the treatment protocol involves

a comparison of placebo vs. treatment response rates in two or more genotype-defined groups. For example a group with no copies of a functional site allele may be compared to a group with two copies of the functional site allele, or a group with no copies may be compared to a group consisting of those with one or two copies of the functional site allele. In this manner different genetic models (dominant, co-dominant, recessive) for the transmission of a treatment response trait can be tested. Alternatively, statistical methods that do not posit a specific genetic model, such as contingency tables, can be used to measure the effects of an allele on treatment response.

In another preferred embodiment, patients in a clinical trial can be grouped (at the end of the trial) according to treatment response, and statistical methods can be used to compare functional site allele (or genotype or haplotype) frequencies in two groups. For example responders can be compared to nonresponders, or patients suffering adverse events can be compared to those not experiencing such effects. Alternatively response data can be treated as a continuous variable and the ability of functional site genotype to predict response can be measured. In a preferred embodiments, patients who exhibit extreme phenotypes are compared with all other patients or with a group of patients who exhibit a divergent extreme phenotype. For example if there is a continuous or semi-continuous measure of treatment response, then the 10% of patients with the most favorable responses could be compared to the 10% with the least favorable, or the patients one standard deviation above the mean score could be compared to the remainder, or to those one standard deviation below the mean score. One useful way to select the threshold for defining a response is to examine the distribution of responses in a placebo group. If the upper end of the range of placebo responses is used as a lower threshold for an "outlier response" then the outlier response group should be almost free of placebo responders. This is a useful threshold because the inclusion of placebo responders in a "true response

group decreases the ability of statistical methods to detect a genetic difference between responders and nonresponders.

The present invention encompasses stratifying patients for human clinical trials. The trial can be a phase I, phase II, phase III or phase IV clinical trial. In other embodiments, invention encompasses practicing the present methods on non-human animals in pre-clinical trials, most preferably non-human mammals, including but not limited to non-human primates, rodents (mice, rats), cats, dogs, pigs, *etc.*

As used herein, the term "stratification" refers to the creation of a distinction between patients on the basis of a characteristic or characteristics of the patient. Generally, in the context of clinical trials, the distinction is used to distinguish responses or effects in different sets of patients distinguished according to the stratification parameters. For the present invention, stratification preferably includes distinction of patient groups based on the presence or absence of particular polymorphisms in one or more functional sites. The stratification may be performed only in the course of analysis or may be used in creation of distinct groups or in other ways.

Other parameters that may be assessed in parallel with functional site-based stratification include gender, race, ethnic or geographic origin (population history) or other demographic factors. In certain embodiments, gene expression profiling (*e.g.*, the use of "signature" gene expression profiles associated with a particular disease) is used in parallel with or prior to functional site-based stratification.

EXAMPLES

OVERVIEW

Libraries of functional sites were constructed. To this end nuclei were treated with DNaseI under conditions which ensured that cutting occurred with

high preference for the DNaseI-hypersensitive sites over the genomic background. DNA was isolated using techniques to reduce the introduction of random breakages during purification of the DNA. This material represented genomic DNA which had been preferentially modified (by introduction of a double-stranded cut) at hypersensitive sites.

The approach for library construction was to create subtractive libraries: a population of Reference DNA molecules (enriched in hypersensitive sequences) were hybridized to a population of an excess of biotinylated Subtractant DNA molecules (depleted in hypersensitive sites). For the two types of libraries described (PS005 and PS008) there were different protocols for the creation of the Subtractant DNA. Following hybridization the DNA duplexes which were non-biotinylated were isolated. These were homoduplexes of Reference DNA and following PCR amplification were cloned to create the libraries. PS005 libraries cloned the PCR fragments directly. To create the PS008 libraries the PCR products were first digested with NlaIII and then cloned into the SphI site of the vector pGEM5z.

A sample of sequences were tested by quantitative Real-time PCR to determine whether or not the genomic location described was indeed a functional ACE. The method used is desrcibed below and an example of a confirmed DNaseI hypersensitive site chosen from a library is shown in Figure 1.

Overview of preparation of Reference DNA

Double stranded breakpoints were repaired by treatment with T4 DNA polymerase to create an homogenous population of ends (being blunted). These were subsequently treated with Taq polymerase in the presence of dNTPs to add an 3' dA overhang. This facilitated the ligation of a biotinyalted adaptor. Following digestion of the genomic DNA with NlaIII the biotinylated fragments were captured and represented a population enriched for those sites associated with the DNA breakpoints. The captured DNA fragments were ligated to an adaptor with a

compatiable NlaIII site. As such the fragments could be be released from the Dynal beads (paramagnetic strepavidin-coated beads used to isolate the biotin-containing DNA) by NotI digestion (a recognition site for which enzyme had been engineered into one of the adaptors). This DNA constituted the Reference DNA and an overview of its production is shown in Figure 2.

Overview of preparation of the PS005 Subtractant DNA

DNA from DNaseI-digested nuclei was digested to completion with NlaIII. On restriction his enzyme generates a four nucleotide 3' overhang which is refractive to digestion by Exonuclease III. The dA overhang, which marks a DNA breakpoint, is a substrate for digestion by Exonuclease III. Hence treatment of the sample with Exonuclease III will convert double stranded fragments containing a dA overhang to single stranded DNA. In an additional step this is degraded by treatment with Mung Bean Nuclease. Recovered DNA is then biotinylated by addition of a biotinylated-ddUTP by the action of Terminal Transferase and chemical modification with Photobiotin. As such the DNA is depleted in hypersensitive sites and biotinylated and is used as the Subtractant DNA for the creation of PS005. A general scheme for the production of Subtractant DNA is shown in Figure 3.

Overview of preparation of the PS008 Subtractant DNA

This procedure is similar as above with the exception that instead of DNA from DNaseI-digested nuclei digestion with NlaIII four aliquots of the DNA are formed and each is digested with either PstI, SphI, NsiI or SacI. These enzymes also produce ends refractive to Exonuclease III digestion. The pools are combined and treated with Exonuclease III and Mung Bean Nuclease and biotinylated as described above to create the PS008 Subtractant DNA. A general scheme for the production of Subtractant DNA is shown in Figure 3.

Subtraction process

The hybridization conditions used are those previously described for the creation of subtractive cDNA libraries (Diachenko et al., 1996. PMID: 8650213).

MATERIALS AND METHODS

Two sets of subtractive genomic libraries (refered to as PS005 and consisting of the libraries PS001 to PS005 inclusive, and PS008 consisting of the libraries PS006 to PS011) enriched in ACEs were created and validated. They shared the same protocol for creation of Reference DNA and the Subtraction step. These steps are described in generic protocols. Library construction differed in the preparation of Subtractive DNA and the final cloning, these protocols are described independently for each library type.

Growth of tissue culture cells and isolation of nuclei

Human K562 chronic myeloid leukemia cell lines were obtained from ATTC and grown in RPMI supplemented with 10% FBS, PenStrep, L-glut, and NaPyr to a cell density of ~5 X 10^5 cells/ml. For nuclei preparation cells were spun at 500 x g for 3 minutes, resuspended in 25ml of PBS, washed once in 25ml PBS and the pellet resuspended briefly in 25ml of buffer A (Reitman *et al.* 1993) containing 1% NP40. Nuclei were then spun at 1000g for 3 min, washed once in buffer A without NP40, and resuspended in at a concentration of 1 X 10^8 nuclei/ml. All nuclei preparation steps were done at 4°C.

Treatment with DNaseI and harvesting genomic DNA

K562 cells were grown to confluence (5 x 10^5 cells per cubit milliliter as assayed by hemocytometer). Nuclei were prepared from a suitable volume (e.g., 100ml) and nuclei were prepared as described (Reitman et al MCB 13:3990). Briefly, Nuclei were resuspended at a concentration of 8 OD/ml with 10 microliters

of 2 U/microliter DNaseI [Sigma] at 37°C for 3 min. The DNA was purified by phenol-chloroform extractions and dialysed into two changes of TE buffer for 2h and overnight. The DNA was repaired in a 100 microliter reaction containing 10 microgram DNA and 6 U T4 DNA polymerase (New England Biolabs) in the manufacturer's recommended buffer and incubated for 15 min at 37°C and then 15 min at 70°C. 1.5 U Taq polymerase (Roche) was added and the incubation continued at 72°C for a further 10 min. The DNA was recovered using a Qiagen DNEasy Clean-up Kit and the DNA eluted in 50 microliter of 10 mM Tris.HCl, pH8.0

Creation of Reference DNA for Subtractive libraries

Nuclei were prepared from K562 cells and resuspended at various amounts (1- 16 U) of DNaseI [Sigma] at 37°C for 3 min as described above. The DNA was purified by phenol-chloroform extractions and purified by extensive dialysis. The DNA was repaired in a 100 µl reaction containing 10 µg DNA and 6 U T4 DNA polymerase (New England Biolabs) in the manufacturer's recommended buffer and incubated for 15 min at 37°C and then 15 min at 70°C. 1.5 U Taq polymerase (Roche) was added and the incubation continued at 72°C for a further 10 min. The DNA was recovered using a Qiagen PCR Clean-up Kit and the DNA eluted in 50 µl of 10 mM Tris.HCl, pH8.0. The DNA was mixed in a 100 µl reaction volume containing 50 pmol of PS003 adapter (created by annealing equimolar amounts of the following oligonucleotides 5'-Biotin-TTA TGC GGC CGC TAT GTG TGC AGT-3'f and 5'-Phosphate-CTG CAC ACA TAG CGG CCG CAT AGG-3', and 40 U T4 DNA ligase (New England Biolabs) in the manufacturer's recommended buffer for 16 h at 4°C.

The reaction was incubated at 65°C for 20 min before the DNA was isopropanol precipitated in the presence of 0.3 M NaOAc and after ethanol washing resuspended in 20 µl TE buffer (10 mM Tris.HCl, 1 mM EDTA, pH8.0). The DNA was digested in a 50 µl reaction volume containing 20 U *Hsp*92 II

(Promega) in the manufacturer's recommended buffer by incubation at 37°C for 2 h, after which a further 20 U of enzyme was added and the incubation continued for 1 h and then heated to 72°C for 15 min. The DNA was captured on M-270 Dynal beads as per manufacturer's instructions. The beads were finally washed in 200 µl of ligation buffer before capture and resuspension in a 100 µl reaction volume containing 50 pmol of Hsp adapter (made by annealing equimolar amounts of the oligonucleotides HSpf 5'-GCG TAC TCC GAC TCG CTA TAG ATC ATG-3' and HSpr 5'-Phosphate-ATC TAT AGC GAG TCG GAG TAC GC-3' supplemented with 6 U T4 DNA ligase (New England Biolabs) in the manufacturer's recommended buffer and incubated at 16°C for 16 h. The reaction was heated to 65°C for 15 min prior to capture of the beads. The beads were washed in 1 x NEB3 buffer (New England Biolabs) and then resuspended in a reaction volume of 100 µl of the same buffer supplemented with 40 U *Not*I (New England Biolabs) and incubated for 37°C for 1 hour with occasional mixing, after which the beads were captured and the supernatant retained. The beads were washed once and the resultant supernatant combined with the first and isopropanol precipitated in the presence of 20 µg glycogen and 0.3 M NaOAc. After ethanol washing the DNA was resuspended in 10 µl of 10 mM Tris.HCl, pH8.0. This DNA is then reserved as the Reference population.

Creation of Subtractant DNA for PS005 Subtractive Libraries

DNA was prepared from K562 nuclei that had been treated with DNaseI as above but with higher concentrations. Ten µg of DNA were digested in four separate reactions with New England Biolabs *Nla*III as per the manufacturer's instructions. The reactions were pooled, heat-inactivated and extracted with phenol before precipitating with ethanol and resuspension in 10 µl of TE. 10 µg digested DNA was digested in a 75 µl reaction volume supplemented with 400 U exonuclease III (Promega) in the manufacturer's recommended buffer for 25°C for 3 min. The reaction was supplemented with 225 µl buffer containing 30 µl Promega

10 x Mung Bean Nuclease buffer and 200 U Mung Bean Nuclease and incubated for a further 25 min. The reaction was stopped by the addition of 30 µl Stop buffer (300 mM Tris.HCl, 50 mM EDTA, pH8.0) and 33 µl 3 M NaOAc. The reaction was extracted with phenol and ethanol precipitated and resuspended in 22 µl TE buffer. The DNA was treated in a 40 µl reaction volume supplemented with 5 mM $CoCl_2$, 25 µM ddUTP-Biotin (Roche) and 25 U Terminal transferase (Roche) for 15 min at 37°C before ethanol precipitation in the presence of 20 µM EDTA and 0.8 M LiCl. The DNA was resuspended in 10 µl TE buffer. The DNA was then additionally labeled with Photobiotin as per the manufacturer's directions. 10 µl of the DNA solution was mixed with 10 µl of photobiotin and after exposure to the UV source for 15 min 30 µl of TE was added and the reaction treated with a Roche G-50 spin column followed by extraction with 2 x water-saturated butanol. The DNA was ethanol precipitated and resuspended in 10 µl water. This DNA was retained as the PS005 Subtractant population.

Creation of Subtractant DNA for PS008 Subtractive Libraries

DNA was prepared from K562 nuclei that had been treated with DNaseI as above but with higher concentrations. Ten µg of DNA were digested in four separate reactions with New England Biolabs *Pst*I, *Sph*I, *Nsi*I or *Sac*I as per the manufacturer's instructions. The reactions were pooled, heat-inactivated and extracted with phenol before precipitating with ethanol and resuspension in 10 µl of TE. 10 µg digested DNA was digested in a 75 µl reaction volume supplemented with 400 U exonuclease III (Promega) in the manufacturer's recommended buffer for 25°C for 3 min. The reaction was supplemented with 225 µl buffer containing 30 µl Promega 10 x Mung Bean Nuclease buffer and 200 U Mung Bean Nuclease and incubated for a further 25 min. The reaction was stopped by the addition of 30 µl Stop buffer (300 mM Tris.HCl, 50 mM EDTA, pH8.0) and 33 µl 3 M NaOAc. The reaction was extracted with phenol and ethanol precipitated and resuspended in 22 µl TE buffer. The DNA was treated in a 40 µl reaction volume supplemented

with 5 mM $CoCl_2$, 25 μM ddUTP-Biotin (Roche) and 25 U Terminal transferase (Roche) for 15 min at 37°C before ethanol precipitation in the presence of 20 μM EDTA and 0.8 M LiCl. The DNA was resuspended in 10 μl TE buffer. The DNA was then additionally labeled with Photobiotin as per the manufacturer's directions. 10 μl of the DNA solution was mixed with 10 μl of photobiotin and after exposure to the UV source for 15 min 30 μl of TE was added and the reaction treated with a Roche G-50 spin column followed by extraction with 2 x water-saturated butanol. The DNA was ethanol precipitated and resuspended in 10 μl water. This DNA was retained as the PS008 Subtractant population.

Subtractive Process

One μl each of the Reference and Subtractant populations (for either PS005 or PS008) were mixed with 5 μl of hybridization buffer (20 mM EPPS, 2 mM EDTA) and 1 μl water and heated to 95°C for 2 min before 2 μl of 5 M NaCl was added and the reaction cooled to 40°C over 1 h. Then the reaction is incubated for 40°C for 16 h. Following this incubation the reaction was captured on Dynal beads for 1 h at 37°C with occasional mixing. Following the capture the supernatant was retained and combined with the supernatant generated by the following wash step. The DNA was ethanol precipitated from these fractions in the presence of 20 μg of glycogen and 0.3 M NaOAc following a phenol extraction. Aliquots of the DNA was used in a 50 ml reaction volume consisting of 1 x Roche Faststart Taq polymerase buffer supplemented with 200 mM dNTPs, 0.5 U Faststart Taq polymerase and 25 pmol each of oligonucleotides PS003f and Hspf. The reaction was performed with the following program: 94C for 5 min. then 20 cycles of 94°C for 15 s, 60°C for 15 s and 72°C for 1 min, followied by 72°C for 5 min. then cooling to 4°C.

Cloning of PS005 libraries

The products from the PCR reaction following substraction of PS005 Subtractant DNA from Reference DNA were cloned into pCRII-TOPO vector

(Invitrogen) as per manufacturer's instructions. The ligation was transformed into DH10B cells.

Cloning of PS008 libraries

The products from the PCR reaction following substraction of PS008 Subtractant DNA from Reference DNA were digested to completion by *Nla*III (New England Biolabs) and cloned into *Sph*I-digested pGEM5z vector (Promega) as per manufacturer's instructions. The ligation was transformed into DH10B cells.

qPCR validation of libraries

The proportion of cloned sequences which were coincident with hypersensitive sites was established by application of a quantitative Real-time PCR assay (McArthur *et al.*, 2001). The reaction is performed following treatment of nuclei with DNaseI the genomic DNA is purified and used as a template in a qPCR reaction. Primers are designed to amplify similarly sized test amplicons and the progress of the amplification is followed by measuring the accumulation of signal from the double stranded DNA-specific fluorescent dye SYBR green. The Real-time PCR system measures this increase in flourescence as a function of cycle number and from the resultant amplification profile calculates the number of cycles of PCR needed to amplify the product above a fixed threshold. Every sample tested is ran synchronously with an undigested template DNA. The number of copies of the test amplicon are normalised with the number of copies of a refernce amplicon- designed in a DNaseI insensitive region of the genome. A digestion profile is generated by calculating the relative loss of a test amplicon across a series of DNaseI digestion conditions and expressing the loss relative to the number copies present in a non-digested sample (set at 100%).

Assessing DNaseI-hypersensitivity by Quantitative Real-time PCR

Analysis was performed on candidate sequences to determine relative copy number loss in a DNaseI-digested sample using quantitative Real-time PCR based on principles previously described (McArthur *et al.*, 2001). A 15μL real-time quantitative PCR reactions was aseembled using 0.9μM forward and reverse primers, 30 ng template DNA (untreated or DnaseI-treated) and master mix composed of 1X FastStart buffer (Roche), 200μM of each dATP, dCTP, dGTP, dTTP, 3mM $MgCl_2$ and FastStart Taq DNA polymerase (0.033 U/μL). The reaction mixture was supplemented with 0.33X SYBR green I stain and 300 nM 6-ROX (Molecular Probes, Eugene, OR) to detect the accumulation of PCR product during amplification and normalize fluorescence intensity, respectively. All qPCR reactions were set up robotically with a Biomek FX (Beckman, Fullerton, CA). Samples were run in triplicate on individual 384-well plates, and thermalcycled with an ABI 7900HT Sequence Detection System (Applied Biosystems, Foster City, CA). Normalized fluorescence data were exported using the ABI SDS software (v2.0). An amplification curve and Nth-order polynomial fit was then computed for each reaction. Cycle threshold (C_t) values were then determined for each curve. The amplification efficiency of a reference amplicon selected from the inactive and DNase-insensitive Rhodopsin locus (3q21-q24) was determined empirically for every reaction plate using a standard dilution series of DNA and the equation $E = 10^{-1/slope}$. We then derived the efficiency of each test amplicon is from the slope of the linear region of the amplification curve. Efficiency corrections were then performed on all test amplicons with respect to the reference amplicon, following which we calculated relative copy number differences using the comparative C_t method (Livak *et al.*, 2001, en TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25:402-8). Melting curve analysis was conducted for each amplicon to discard those yielding multiple products. Efficiency-corrected C_t values were then used to

compute a relative copy number ratio by applying the formula $2^{-\Delta\Delta Ct}$ or $2^{-[\text{treated (target – reference)} - \text{calibrator (target-reference)}]}$. Relative DNaseI sensitivity ratios (=relative copy ratios) were thus obtained. Ratios < 1 are indicative of relative copy loss due to preferential cleavage of chromatin by DNaseI.

REFERENCES

Adams, C. C. & Workman, J. L. (1995). Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. Mol Cell Biol 15, 1405-1421.

Aladjem, M. I., Rodewald, L. W., Kolman, J. L. & Wahl, G. M. (1998). Genetic dissection of a mammalian replicator in the human beta-globin locus. Science 281, 1005-9.

Almer, A., and Horz, W. (1986) Nuclease hypersensitive regions with adjacent positioned nucleosomes mark the gene boundaries of the PHO5/PHO3 locus in yeast. Embo J 5: 2681-2687.

Altschul, S.F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). A basic local alignment search tool. J Mol Biol 215:403-10.

Anderson, D.C., Li, W., Payan, D. G., & Noble, W. S. (2003). A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. J Proteome Res. To appear.

Arima, M., Toyama, H., Ichii, H., Kojima, S., Okada, S., Hatano, M., Cheng, G., Kubo, M., Fukuda, T., and Tokuhisa, T. (2002) A putative silencer element in the IL-5 gene recognized by Bcl6. J Immunol 169: 829-836.

Aronow, B.J., Ebert, C.A., Valerius, M.T., Potter, S.S., Wiginton, D.A., Witte, D.P., Hutton, J.J.(1995). Dissecting a locus control region: facilitation of enhancer function by extended enhancer-flanking sequences. Mol Cell Biol 15:1123-1135.

Aronow, B.J., Silbiger, R.N., Dusing, M.R., Stock, J.L., Yager, K.L., Potter, S.S., Hutton, J.J., Wiginton, D.A. (1992). Functional analysis of the human adenosine deaminase gene thymic regulatory region and its ability to generate position-independent transgene expression. Mol Cell Biol 12(9):4170-4185.

Bailey, T. L. & Elkan, C. P. (1994). Fitting a mixture model by expectation-maximization to discover motifs in biopolymers. In R. Altman, D. Brutlag, P. Karp, R. Lathrop, and D. Searls, editors, Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pages 28-36. AAAI Press.

Bailey, T. L. & Noble, W. S. Searching for statistically significant regulatory modules. Submitted.

Beasley, E.M. , Myers, R.M., Cox, D.R., Lazzeroni, L.C. (1999). Statistical Refinement of Primer Design Parameters. In: PCR Applications. Innis, M.A, Gelfand, D.H., and Sninsky, J. J., eds. Academic Press, San Diego, pp 55-71.

Bender, M.A., Mehaffey, M.G., Telling, A., Hug, B., Ley, T.J., Groudine, M., Fiering, S. (2000). Independent formation of DnaseI hypersensitive sites in the murine beta-globin locus control region. Blood. 95:3600-3604.

Berman, B. P., Nibu, Y., Pfeifer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M. & Eisen, M. B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci USA, 99:757-762.

Blanchette, M. & Tompa, M. (2002). Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res, 12:739-748.

Bock, J. R. & Gough, D. A. (2001). Predicting protein-protein interactions from primary structure. Bioinformatics, 17:455-460.

Boyd, S., El Ghaoui, L., Feron, E. & Balakrishnan, V. (June 1994) Linear Matrix Inequalities in System and Control Theory, volume 15 of Studies in Applied Mathematics. SIAM, Philadelphia, PA.

Boyes, J., Felsenfeld, G. (1996). Tissue-specific factors additively increase the probability of the all-or-none formation of a hypersensitive site. EMBO J 15:2496-2507.

Bresnick et al., (1992) The transcriptionally-active MMTV promoter is depleted of histone H1. Nucl. Acids Res. 20:273-278

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C., Furey, Jr., T. S., Ares, M. & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data using support vector machines. Proc Natl Acad Sci USA, 97:262-267.

Brownell, J. E. & Allis, C. D. (1996). Special HATs for special occasions: linking histone acetylation to chromatin assembly and gene activation. Curr Opin Genet Dev 6, 176-84.

Burgess-Beusse, B., Farrell, C., Gaszner, M., Litt, M., Mutskov, V., Recillas-Targa, F., Simpson, M., West, A. & Felsenfeld, G. (2002). The insulation of genes from external enhancers and silencing chromatin. Proc Natl Acad Sci U S A.

Bussemaker, H.J., Li, H. & Siggia, E. D. (2001). Regulatory element detection using correlation with expression. Nature Genet, 27:167-171.

Butler, J.E., and Kadonaga, J.T. (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev 16: 2583-2592.

Chambers, J.M., Cleveland, W.S., Tukey, P.A. (1983). Graphical Methods for Data Analysis. Belmont, CA: Wadsworth.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. & Herskowitz, I. (1998) The transcriptional program of sporulation in budding yeast. Science, 282:699-705.

Cristianini, N. &Shawe-Taylor, J. (1997). An Introduction to Support Vector Machines. Cambridge UP, 2000.

Crowley, E. M., Roeder, K. & Bina, M. . A statistical model for locating regulatory regions in genomic DNA. J Mol Biol, 268:8-14.

de Moura Gallo, C.V., Vassetzky, Y.S., Huesca, M., and Scherrer, K. (1992) A transcription-dependent DNase I-hypersensitive site in a far upstream segment of the chicken alpha-globin gene domain coincides with a matrix attachment region. Biochem Biophys Res Commun 184: 1226-1234.

Ding, C. & Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics, 17:349-358.

Duret, L. & Bucher, P. (1997). Searching for regulatory elements in human noncoding sequences. Current Opin Struct Biol, 7:399-405.

Elgin, S. C. (1981). DNAase I-hypersensitive sites of chromatin. Cell 27, 413-5.

Elgin, S. C. (1984). Anatomy of hypersensitive sites. Nature 309, 213-4.

Elgin, S.C. (1988) The formation and function of DNase I hypersensitive sites in the process of gene activation. J Biol Chem 263: 19259-19262.

Eskin, E. & Pevzner, P. A. (2002). Finding composite regulatory patterns in DNA sequences. Bioinformatics, 18:S354-S363.

Eskin, E., Keich, U., Gelfand, M.S. & Pevzner, P.A. (2003). Genome-wide analysis of bacterial promoter regions. In Proceedings of the Pacific Symposium on Biocomputing, pages 29-40, New Jersey. World Scientific.

Farkas, G., Leibovitch, B.A., and Elgin, S.C. (2000) Chromatin organization and transcriptional control of gene expression in Drosophila. Gene 253: 117-136.

Felsenfeld, G. & Groudine, M. (2003). Controlling the double helix. Nature 421, 448-53.

Felsenfeld, G. (1992). Chromatin as an essential part of the transcriptional mechanism. Nature 355, 219-24.

Felsenfeld, G. (1996). Chromatin unfolds. Cell 86, 13-9.

Fiering, S., Whitelaw, E., Martin, D.I. (2000). To be or not to be active: the stochastic nature of enhancer action. Bioessays 22:381-387.

Forrester, W.C., Thompson, C., Elder, J.T., Groudine, M. (1986). A developmentally stable chromatin structure in the human beta-globin gene cluster. Proc Natl Acad Sci U S A. 83:1359-1363.

Forsberg, E. C., Downs, K. M. & Bresnick, E. H. (2000). Direct interaction of NF-E2 with hypersensitive site 2 of the beta-globin locus control region in living cells. Blood 96, 334-9.

Fraser et al., (1990) DNAseI hypersensitive sites 1,2 and 3 of the human beta-globin dominant control region direct position-independnet expression. Nuc. Acids Res. 18:3503-3508

Frech, K., Danescu-Mayer, J. & Werner, T. (1997) A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. J Mol Biol, 270:674-687.

Friedman, H.S., Burger, P.C., Bigner, S.H., Trojanowski, J.Q., Brodeur, G.M., He, X.M., Wikstrand, C.J., Kurtzberg, J., Berens, M.E., Halperin, E.C., and et al. (1988) Phenotypic and genotypic analysis of a human medulloblastoma cell line and transplantable xenograft (D341 Med) demonstrating amplification of c-myc. Am J Pathol 130: 472-484.

Friedman, H.S., Burger, P.C., Bigner, S.H., Trojanowski, J.Q., Wikstrand, C.J., Halperin, E.C., and Bigner, D.D. (1985) Establishment and characterization of the human medulloblastoma cell line and transplantable xenograft D283 Med. J Neuropathol Exp Neurol 44: 592-605.

Frith, M. C., Hansen, U. & Weng, Z. (2001). Detection of cis-element clusters in higher eukaryotic DNA. Bioinformatics, 17:878-889.

Frith, M. C., Spouge, J.L., Hansen, U. & Weng, Z. (2002). Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. Nucleic Acids Res, 30:3214-3224.

Fürbass, R., Said, H. M., Schwerin, M. & Vanselow, J. (2001). Chromatin structure of the bovine Cyp19 promoter 1.1. DNaseI hypersensitive sites and DNA hypomethylation correlate with placental expression. Eur J Biochem 268, 1222-7.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M. & Haussler, D. (2001). Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16:906-914.

Giardine, B., Elnitski, L., Riemer, C., Makalowska, I., Schwartz, S., Miller, W., Hardison, R.C. (2003). GALA, a database for genomic sequence alignments and annotations. Genome Res 13:732-741.

Gill, G. (2001) Regulation of the initiation of eukaryotic transcription. Essays Biochem 37: 33-43.

Ginzinger, D. G. (2002). Gene quantification using real-time quantitative PCR: an emerging technology hits the mainstream. Exp Hematol 30, 503-12.

Goodwin, A. J., McInerney, J. M., Glander, M. A., Pomerantz, O. & Lowrey, C. H. (2001). In vivo formation of a human beta-globin locus control region core element requires binding sites for multiple factors including GATA-1, NF-E2, erythroid Kruppel-like factor, and Sp1. J Biol Chem 276, 26883-92.

Gourdon, G., Sharpe, J.A., Wells, D., Wood, W.G., Higgs, D.R. (1994). Analysis of a 70 kb segment of DNA containing the human zeta and alpha-globin genes linked to their regulatory element (HS-40) in transgenic mice. Nucleic Acids Res 22:4139-4147.

Gross, D.S. and Garrard, W. T. (1988) Nuclease hypersensitive sites in chromatin. Annu. Rev. Biochem. 57:159-197.

Groudine, M., Kohwi-Shigematsu, T., Gelinas, R., Stamatoyannopoulos, G., Papayannopoulou, T. (1983). Human fetal to adult hemoglobin switching: changes in chromatin structure of the beta-globin gene locus. Proc Natl Acad Sci U S A. 80:7551-7555.

Groudine, M., Weintraub, H. (1982). Propagation of globin DNAase I-hypersensitive sites in absence of factors required for induction: a possible mechanism for determination. Cell 30:131-139.

Grunstein, M. (1997). Histone acetylation in chromatin structure and transcription. Nature. 389:349-352.

Gumucio, D.L., Shelton, D.A., Zhu, W., Millinoff, D., Gray, T., Bock, J.H., Slightom, J.L., Goodman, M. (1996). Evolutionary strategies for the elucidation of cis and trans factors that regulate the developmental switching programs of the beta-like globin genes. Mol Phylogenet Evol 5:18-32.

Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. (2001). Gene selection for cancer classification using support vector machines. Machine Learning.

Hache, R.J., and Deeley, R.G. (1988) Organization, sequence and nuclease hypersensitivity of repetitive elements flanking the chicken apoVLDLII gene: extended sequence similarity to elements flanking the chicken vitellogenin gene. Nucleic Acids Res 16: 97-113.

Hardison, R., Miller, W. (1993). Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters. Mol Biol Evol. 10:73-102.

Hardison, R., Slightom, J.L., Gumucio, D.L., Goodman, M., Stojanovic, N., Miller, W. (1997). Locus control regions of mammalian beta-globin gene clusters: combining phylogenetic analyses and experimental results to gain functional insights. Gene 205:73-94.

Hardison, R.C. (2001). New views of evolution and regulation of vertebrate beta-like globin gene clusters from an orphaned gene in marsupials. Proc Natl Acad Sci U S A Feb 98:1327-1329.

He, X.M., Wikstrand, C.J., Friedman, H.S., Bigner, S.H., Pleasure, S., Trojanowski, J.Q., and Bigner, D.D. (1991) Differentiation characteristics of newly established medulloblastoma cell lines (D384 Med, D425 Med, and D458 Med) and their transplantable xenografts. Lab Invest 64: 833-843.

Hertz, G.Z. & Stormo, G.D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics, 15:563-577.

Higgs, D.R., Wood, W.G., Jarman, A.P., Sharpe, J., Lida, J., Pretorius, I.M., Ayyub, H. (1990). A major positive regulatory region located far upstream of the human alpha-globin gene locus. Genes Dev 4:1588-1601.

Holloway, M.P., and La Gamma, E.F. (1992) Tissue-specific DNaseI hypersensitivity regions are located in the 5'-region of the rat preproenkephalin gene. J Biol Chem 267: 19819-19823.

Holmes, I. & Bruno, W. J. (2000). Finding regulatory elements using joint likelihoods for sequence and expression profile data. In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, pages 202-210.

Jaakkola, T., Diekhans, M. & Haussler, D. (1999). Using the Fisher kernel method to detect remote protein homologies. In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, pages 149-158, Menlo Park, CA. AAAI Press.

Jarman, A.P., Wood, W.G., Sharpe, J.A., Gourdon, G., Ayyub, H., Higgs, D.R. (1991). Characterization of the major regulatory element upstream of the human alpha-globin gene cluster. Mol Cell Biol 11:4679-4689.

Kadonaga, J. T. (1998). Eukaryotic transcription: an interlaced network of transcription factors and chromatin-modifying machines. Cell 92, 307-13.

Kamvysselis, M., Patterson, N., Birren, B., Berger, B. & Lander, E. (2003). Whole-genome comparative annotation and regulatory motif discovery in multiple yeast species. In Proceedings of the Seventh Annual International Conference on Computational Molecular Biology, pages 157-166.

Keene, M.A., Corces, V., Lowenhaupt, K., and Elgin, S.C. (1981) DNase I hypersensitive sites in Drosophila chromatin occur at the 5' ends of regions of transcription. Proc Natl Acad Sci U S A 78: 143-146.

Kent, W.J., Sugnet, C. W., Furey, T.S., Roskin, K. M., Pringle, T.H., Zahler, A. M. & Haussler, D. (2002). Human genome browser at UCSC. Genome Res, 12:996-1006.

Kieffer et al. (2002) Identification of a candidate regulatory region in the human CD8 gene complex by co-localization of DNAseI hypersensitive sites and matrix attachment regions which bind SATB1 and GATA-3. J. Immunol. 168:3915-3922

Kim, A. R. & Murray, V. (2001). Chromatin structure at the 3'-boundary of the human beta-globin locus control region hypersensitive site-2. Int J Biochem Cell Biol 33, 1183-92.

Kingston, R. E., Bunker, C. A. & Imbalzano, A. N. (1996). Repression and activation by multiprotein complexes that alter chromatin structure. Genes Dev 10, 905-20.

Kondrakhin, Y.V., Kel, E., Kolchanov, N.A., Romashchenko, A.G. & Milanesi, L. (1995). Eukaryotic promoter recognition by binding sites for transcription factors. Comput App Biosci, 11:477-488.

Kong et al. (1997) Transcription of the HS2 enhancer toward a cis-linked gene is independent of the orientation, position and distance of the enhancer relative to the gene. Mol. Cell. Biol. 17:3955-3965

Koropatnick, J., and Duerksen, J.D. (1987) Nuclease sensitivity of alpha-fetoprotein, metallothionein-1, and immunoglobulin gene sequences in mouse during development. Dev Biol 122: 1-10.

Krivan, W. & Wasserman, W. (2001). A predictive model for regulatory sequences directing liver-specific transcription. Genome Res, 11:1559-1566.

Kunnath and Locker (1985) DNaseI sensitivity of the rat albumen and ά-fetoprotein genes. Nucl. Acids Res. 13:115-129

Lanckriet, G.R.G., Cristianini, N., Bartlett, P., El Ghaoui, L. & Jordan, M.I. (2002). Learning the kernel matrix with semi-definite programming. Technical Report CSD-02-1206, University of California, Berkeley.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J. S. Neuwald, A.F. & Wootton, J.C. (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. Science, 262(5131):208-214.

Leslie, C., Eskin, E. & Noble, W.S. (2002) The spectrum kernel: A string kernel for SVM protein classification. In Proceedings of the Pacific Symposium on Biocomputing. To appear.

Levy, S. & Hannenhalli, S. (2002). Identification of transcription factor binding sites in the human genome sequence. Mammalian Genome, 13:510-514.

Levy-Wilson, B., Paulweber, B., Antes, T. J., Goodart, S. A. & Lee, S. Y. (2000). An open chromatin structure in a liver-specific enhancer that confers high level expression to human apolipoprotein b transgenes in mice. Mol Cell Biol Res Commun 4, 206-11.

Li, Q. & Stamatoyannopoulos, G. (1994). Hypersensitive site 5 of the human beta locus control region functions as a chromatin insulator. Blood 84, 1399-401.

Li, Q., Harju, S., and Peterson, K.R. (1999) Locus control regions: coming of age at a decade plus. Trends Genet 15: 403-408.

Li, Q., Peterson, K. R., Fang, X. & Stamatoyannopoulos, G. (2002a). Locus control regions. Blood 100, 3077-86.

Li, Q., Peterson, K.R., and Stamatoyannopoulos, G. (1998) Developmental control of epsilon- and gamma-globin genes. Ann N Y Acad Sci 850: 10-17.

Li, Q., Zhang, M., Duan, Z., and Stamatoyannopoulos, G. (1999) Structural analysis and mapping of DNase I hypersensitivity of HS5 of the Ūglobin locus control region. Genomics 61:183-193.

Li, Q., Zhang, M., Duan, Z., Stamatoyannopoulos, G. (1999). Structural analysis and mapping of DNase I hypersensitivity of HS5 of the beta-globin locus control region. Genomics 61:183-193.

Li, Q., Zhang, M., Han, H., Rohde, A. & Stamatoyannopoulos, G. (2002b). Evidence that DNase I hypersensitive site 5 of the human beta-globin locus control region functions as a chromosomal insulator in transgenic mice. Nucleic Acids Res 30, 2484-91.

Liao, L. & Noble, W.S. (2002). Combining pairwise sequence similarity and support vector machines for remote protein homology detection. In Proceedings of the Sixth Annual International Conference on Computational Mol Biol, pages 225-232.

Liu et al. (2002) Chromatin structural analyses of the mouse Igκ gene locus reveal new hypersensitive sites specifying a transcriptional silencer and enhancer. J. Biol. Chem. 277:32640-32649

Lowrey, C. H., Bodine, D. M. & Nienhuis, A. W. (1992). Mechanism of DNase I hypersensitive site formation within the human globin locus control region. Proc Natl Acad Sci U S A 89, 1143-7.

Lu et al. (1995) The role of a positioned nucleosome at the Drosophila melanogaster hsp26 promoter. EMBO J. 14:4738-4746

Markstein, M., Markstein, P., Markstein, V. & Levine, M.S. (2002). Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. Proc Natl Acad Sci USA, 99:763-768.

Mautner, J., Joos, S., Werner, T., Eick, D., Bornkamm, G.W., Polack, A. (1995). Identification of two enhancer elements downstream of the human c-myc gene. Nucleic Acids Res 23:72-80.

McArthur, M., Gerum, S. & Stamatoyannopoulos, G. (2001). Quantification of DNaseI-sensitivity by real-time PCR: quantitative analysis of DNaseI-hypersensitivity of the mouse beta-globin LCR. J Mol Biol 313, 27-34.

McCue, L., Thompson, W., Carmack, C., Ryan, M.P., Liu, J.S., Derbyshire, V. & Lawrence, C.E. (2001). Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. Nucleic Acids Res, 29:774-782.

Melcher, R., Koehler, S., Steinlein, C., Schmid, M., Mueller, C.R., Luehrs, H., Menzel, T., Scheppach, W., Moerk, H., Scheurlen, M., Koehrle, J., and Al-Taie, O. (2002) Spectral karyotype analysis of colon cancer cell lines of the tumor suppressor and mutator pathway. Cytogenet Genome Res 98: 22-28.

Mouse Genome Sequencing Consortium. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature, 420:520-562.

Mukherjee, S., Tamayo, P., Slonim, D., Verri, A., Golub, T., Mesirov, J. & T. Poggio. (1999). Support vector machine classification of microarray data. Technical Report AI Memo 1677, Massachusetts Institute of Technology.

Nesterov, Y. & Nemirovsky, A. (1994). Interior-point polynomial methods in convex programming: Theory and applications, volume 13 of Studies in Applied Mathematics. SIAM, Philadelphia, PA.

Orkin, S.H. (1995). Regulation of globin gene expression in erythroid cells. Eur J Biochem. 231:271-281.

Ortiz, B.D., Cado, D., and Winoto, A. (1999) A new element within the T-cell receptor alpha locus required for tissue-specific locus control region activity. Mol Cell Biol 19: 1901-1909.

Pavlidis, P., Furey, T. S., Liberto, M., Haussler, D. & Grundy, W.N. (2001). Promoter region-based classification of genes. In Proceedings of the Pacific Symposium on Biocomputing, pages 151-163.

Philipsen, S., Pruzina, S. & Grosveld, F. (1993). The minimal requirements for activity in transgenic mice of hypersensitive site 3 of the beta globin locus control region. EMBO J 12, 1077-85.

Prestridge, D. (1995). Predicting Pol II promoter sequences using transcription factor binding sites. J Mol Biol, 249: 923-932.

Pullner, A., Mautner, J., Albert, T., and Eick, D. (1996) Nucleosomal structure of active and inactive c-myc genes. J Biol Chem 271: 31452-31457.

Radomska, H.S., Satterthwaite, A.B., Burn, T.C., Oliff, I.A., Huettner, C.S., and Tenen, D.G. (1998) Multiple control elements are required for expression of the human CD34 gene. Gene 222: 305-318.

Reitman, M., Lee, E., Westphal, H. & Felsenfeld, G. (1993). An enhancer/locus control region is not sufficient to open chromatin. Mol Cell Biol 13, 3990-8.

Riemer, C., ElSherbini, A., Stojanovic, N., Schwartz, S., Kwitkin, P.B., Miller, W., Hardison, R. (1998). A database of experimental results on globin gene expression. Genomics 53:325-337.

Rousseeuw, P.J., van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points. J Am Stat Assoc 85:633-639.

Rozen, S., Skaletsky, H.J. (2000). Primer3 on the WWW for general users and for biologist programmers. In: Bioinformatics Methods and Protocols: Methods in Molecular Biology. Krawetz, S., Misener, S., eds. Humana Press, Totowa, NJ, pp 365-386.

Sambrook, J., and Russell, D.W. (2001) Molecular cloning: A laboratory manual. Cold Spring Harbor Laboratory Press, New York.

Schölkopf, B., Burges, C. J. C. & Smola, A.J. editors. (1999). Advances in Kernel Methods: Support Vector Learning. MIT Press, Cambridge, MA.

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98, 503-17.

Stamatoyannopoulos, G. and Grosveld, F. Hemoglobin Switching. In Stamatoyannopoulos, G, Majerus, P., Perlmutter, R., Varmus, H. (2001). The molecular basis of blood diseases. 3rd edit, W.B. Saunders, Philadelphia.

Stamatoyannopoulos, J.A., Clegg, C.H. & Li, Q. (1997). Sheltering of gamma-globin expression from position effects requires both an upstream locus control region and a regulatory element 3' to the A gamma-globin gene. Mol Cell Biol 17, 240-7.

Stamatoyannopoulos, J.A., Goodwin, A., Joyce, T. & Lowrey, C.H. (1995). NF-E2 and GATA binding motifs are required for the formation of DNase I hypersensitive site 4 of the human beta-globin locus control region. EMBO J 14, 106-16.

Stratling, W.H., Dolle, A., and Sippel, A.E. (1986) Chromatin structure of the chicken lysozyme gene domain as determined by chromatin fractionation and micrococcal nuclease digestion. Biochemistry 25: 495-502.

Stratling, W.H., Dolle, A., and Sippel, A.E. (1986) Chromatin structure of the chicken lysozyme gene domain as determined by chromatin fractionation and micrococcal nuclease digestion. Biochemistry 25: 495-502.

Struhl, K. (1998). Histone acetylation and transcriptional regulatory mechanisms. Genes Dev 12, 599-606.

Struhl, K. (2001) Gene regulation. A paradigm for precision. Science 293:1054-1055.

Svaren, J., Horz, W. (1996). Regulation of gene expression by nucleosomes. Curr Opin Genet Dev. 6:164-170.

Tagle, D.A., Koop, B.F., Goodman, M., Slightom, J.L., Hess, D.L. & Jones, R.T. (1988). Embryonic ² and fl globin genes of a prosimian primate (Galago crassicaudatus) nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol, 203:439-455.

Tanaka, H., Zhao, Y., Wu, D., and Hersh, L.B. (1998) The use of DNase I hypersensitivity site mapping to identify regulatory regions of the human cholinergic gene locus. J Neurochem 70: 1799-1808.

Thomas, J. O. (1984). The higher order structure of chromatin and histone H1. J Cell Sci Suppl 1, 1-20.

Tolhuis, B., Palstra, R. J., Splinter, E., Grosveld, F. & de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell 10, 1453-65.

Tsukiyama, T. & Wu, C. (1997). Chromatin remodeling and transcription. Curr Opin Genet Dev 7, 182-91.

Tuan, D., Solomon, W., Li, Q., London, I.M. (1985). The "beta-like-globin" gene domain in human erythroid cells. Proc Natl Acad Sci U S A. 82:6384-6388.

Urnov et al. (2002) A DNaseI hypersensitive site flanks an origin of DNA replication and amplification on Sciara. Chromosoma 111:291-303

van Drunen, C. M., Sewalt, R. G., Oosterling, R. W., Weisbeek, P. J., Smeekens, S. C. & van Driel, R. (1999). A bipartite sequence element associated with matrix/scaffold attachment regions. Nucleic Acids Res 27, 2924-30.

van Helden, J., Rios, A. F. & Collado-Vides, J. (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res, 28:1808-1818.

Vandenberghe, L. & Boyd, S. (1996). Semidefinite programming. SIAM Review, 38:49-95.

Vapnik, V.N. (1998). Statistical Learning Theory. Adaptive and learning systems for signal processing, communications, and control. Wiley, New York.

Vert, J-P. & Kanehisa, M. (2003). Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, Advances in Neural Information Processing Systems 15. MIT Press.

Wallrath, L.L., Lu, Q., Granok, H., Elgin, S.C. (1994). Architectural variations of inducible eukaryotic promoters: preset and remodeling chromatin structures. Bioessays 16:165-170.

Wasserman, W.W. & Fickett, J.W. (1998). Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol, 278:167-181.

Weintraub, H. & Groudine, M. (1976). Chromosomal subunits in active genes have an altered conformation. Science 193, 848-56.

Weintraub, H., and Groudine, M. (1976) Chromosome subunits in active genes have an altered conformation. Science 193:848-856.

Wen, S. C., Roder, K., Hu, K. Y., Rombel, I., Gavva, N. R., Daftari, P., Kuo, Y. Y., Wang, C. & Shen, C. K. (2000). Loading of DNA-binding factors to an erythroid enhancer. Mol Cell Biol 20, 1993-2003.

Wolffe, A. P., Wong, J. & Pruss, D. (1997). Activators and repressors: making use of chromatin to regulate transcription. Genes Cells 2, 291-302.

Xing, E.P., Jordan, M.I., Karp, R.M. & Russell, S. (2003). A hierarchical bayesian markovian model for motifs in biopolymer sequences. In S. Becker, S. Thrun, and K. Obermayer, editors, Advances in Neural Information Processing Systems, Cambridge, MA. MIT Press.

Youn, B. S., Lim, C. L., Shin, M. K., Hill, J. M. & Kwon, B. S. (2002). An intronic silencer of the mouse perforin gene. Mol Cells 13, 61-8.

Zhang et al. (2002) Genomic DNA breakpoints in AML1/RUNX1 and ETO cluster with topoisomerase II DNA cleavage and DNase I hypersensitive sites in t(8:21) leukemia. Proc. Natl. Acad. Sci. USA 99:3070-3075

Zien, A., Rätch, G., Mika, S., Schölkopf, B., Lengauer, T. & Müller, K.-R. (2000). Engineering support vector machine kernels that recognize translation initiation sites. Bioinformatics, 16(9):799-807.

ABOUHEIF, E. H., and G. A. WRAY. 2002. The developmental genetic basis for the evolution of wing polyphenism in ants. Science 297:249-252.

ABRAHAM, I., and W. W. DOANE. 1978. Genetic regulation of tissue-specific expression of amylase structural genes in Drosophila melanogaster. Proc Natl Acad Sci USA 75:4446-4450.

ABU-SHAAR, M., H. D. RYOO, and R. S. MANN. 1999. Control of the nuclear localization of Extradenticle by competing nuclear import and export signals. Gen & Dev 13:935- 945.

ABZHANOV, A., and T. C. KAUFMAN. 2000. Crustacean (malacostracan) Hox genes and the evolution of the arthropod trunk. Development 127:2239-2248.

AKASHI, H. 2001. Gene expression and molecular evolution. Curr Op Gen & Dev 11:660-666.

ALBERTS, B., A. JOHNSON, J. LEWIS, M. RAFF, K. ROBERTS, and P. WALTER. 2002. The molecular biology of the cell. Garland Publishing, New York.

ALLENDORF, F. W., K. L. KNUDSEN, and R. F. LEARY. 1983. Adaptive significance of differences in the tissue-specific expression of a phosphoglucomutase gene in rainbow trout. Proc Natl Acad Sci USA 80:1397-1400.

ALLENDORF, F. W., K. L. KNUDSEN, and S. R. PHELPS. 1982. Identification of a gene regulating the tissue expression of a phosphoglucomutase locus in rainbow trout.

Genetics 102:259-268.

ANDERSSON, C. R., E. O. JENSEN, D. J. LLEWELLYN, E. S. DENNIS, and W. J. PEACOCK. 1996. A new hemoglobin gene from soybean: A role for hemoglobin in all plants. Proc Natl Acad Sci USA 93:5682-5687.

ANGERER, L. M., D. W. OLEKSYN, A. M. LEVINE, X. LI, W. H. KLEIN, and R. C. ANGERER. 2001. Sea urchin goosecoid function links fate specification along the animal-vegetal and oral-aboral embryonic axes. Development 128:4393-4404.

ANISIMOV, S. V., M. V. VOLKOVA, L. V. LENSKAYA, V. K. KHAVINSON, D. V. SOLOVIEVA, and E. I. SCHWARTZ. 2001. Age-associated accumulation of the apolipoprotein C-III gene T-455C polymorphism C allele in a Russian population. J Gerontol A Biol Sci Med Sci 56:B27-32.

APARICIO, S., A. MORRISON, A. GOULD, J. GILTHORPE, C. CHAUDHURI, P. RIGBY, R. KRUMLAUF, and S. BRENNER. 1995. Detecting conserved regulatory elements with the model genome of the Japanese pufferfish, Fugu rubipes. Proc Natl Acad Sci USA 92:1684-1688.

ARBEITMAN, M. N., E. E. FURLONG, F. IMAM et al. (and 7 co-authors). 2002. Gene expression during the life cycle of Drosophila melanogaster. Science 297:2270-2275.

ARNONE, M. I., and E. H. DAVIDSON. 1997. The hardwiring of development: Organization and function of genomic regulatory systems. Development 124:1851-1864.

ARNONE, M. I., E. L. MARTIN, and E. H. DAVIDSON. 1998. Cis-regulation downstream of cell type specification: A single compact element controls the complex expression of the Cylla gene in sea urchin embryos. Development 125:1381-1395.

ATCHISON, M. L. 1988. Enhancers: mechanisms of action and cell specificity. Annu Rev Cell Biol 4:127-153.

AVEROF, M., and N. H. PATEL. 1997. Crustacean appendage evolution associated with

changes in Hox gene expression. Nature 388:682-686.

AVILA, S., M. C. CASERO, R. FERNANDEZ-CANTON, and L. SASTRE. 2002. Transactivation domains are not functionally conserved between vertebrate and invertebrate serum response factors. Eur J Biochem 269:3669-3677.

BABICH, V., N. AKSENOV, V. ALEXEENKO, S. L. OEI, G. BUCHLOW, and N. TOMILIN. 1999. Association of some potential hormone response elements in human genes with the Alu family repeats. Gene 239:341-349.

BAMSHAD, M. J., S. MUMMIDI, E. GONZALEZ et al. (and 8 co-authors). 2002. A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. Proc Natl Acad Sci USA 99:10539-10544.

BARRIER, M., R. H. ROBICHAUX, and M. D. PURUGGANAN. 2001. Accelerated regulatory gene evolution in an adaptive radiation. Proc Natl Acad Sci USA 98:10208-10213.

BECKERS, J., and D. DUBOULE. 1998. Genetic analysis of a conserved sequence in the HoxD complex: Regulatory redundancy or limitations of the transgenic approach? Dev Dyn 213:1-11.

BELDADE, P., P. M. BRAKEFIELD, and A. D. LONG. 2002. Contribution of Distal-less to quantitative variation in butterfly eyespots. Nature 415:315-318.

BELL, A. C., and G. FELSENFELD. 1999. Stopped at the border: Boundaries and insulators. Curr Op Gen & Dev 9:191-198.

BELL, S. D., and S. P. JACKSON. 1998. Transcription in Archaea. Cold Spring Harbor Symp Quant Biol 63:41-51.

BELTING, H.-G., C. S. SHASHIKANT, and F. H. RUDDLE. 1998. Modification of expression and cis-regulaton of Hoxc8 in the evolution of diverged axial morphology. Proc Natl Acad Sci USA 95:2355-2360.

BENBROOK, D. M., and N. C. JONES. 1990. Heterodimer formation between CREB and JUN proteins. Oncogene 5:295-302.

BENDER, W., M. AKAM, F. KARCH, P. A. BEACHY, M. PEIFER, P. SPIERER, E. B. LEWIS, and D.
S. HOGNESS. 1983. Molecular-genetics of the Bithorax complex in Drosophila melanogaster. Science 221:23-29.

BENECKE, A., C. GAUDON, and H. GRONEMEYER. 2001. Transcriptional integration of 89
hormone and metabolic signals by nuclear receptors. Pp. 167-214 in J. Locker, ed. Transcription Factors. Academic Press, Inc., San Diego.

BENEZRA, R., R. L. DAVIS, D. LOCKSHON, D. L. TURNER, and H. WEINTRAUB. 1990. The
protein Id: a negative regulator of helix-loop-helix DNA-binding proteins. Cell 61:49-59.

BERGMAN, C. M., and M. KREITMAN. 2001. Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. Genome Res 11:1335-1345.

BERMAN, B. P., Y. NIBU, B. D. PFEIFFER, P. TOMANCAK, S. E. CELNICK, M. LEVINE, G. M.
RUBIN, and M. B. EISEN. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci USA 99:757-767.

BERNSTEIN, B. E., J. K. TONG, and S. L. SCHREIBER. 2000. Genome-wide studies of histone
deacetylase function in yeast. Proc Natl Acad Sci USA 97:13708-13713.

BERTHELSEN, J., V. ZAPPAVIGNA, E. FERETTI, F. MAVILIO, and F. BLASI. 1998. The novel
homeoprotein Prep1 modulates Pbx-Hox protein cooperativity. EMBO J 17:1434-1445.

BETRAN, E., K. THORNTON, and M. LONG. 2002. Retroposed new genes out of the X in Drosophila. Genome Res 12:1854-1859.

BHARATHAN, G., T. E. GOLIBER, C. MOORE, S. KESSLER, T. PHAM, and N. R. SINHA. 2002.
Homologies in leaf form inferred from KNOXI gene expression during development. Science 296:1858-1860.

BHARATHAN, G., B.-J. JANSSEN, E. A. KELLOGG, and N. SINHA. 1997. Did homeodomain proteins duplicate before the origin of angiosperms, fungi, and metazoa? Proc Natl Acad Sci USA 94:13749-13753.

BIGGIN, M. D., and W. MCGINNIS. 1997. Regulation of segmentation and segmental identity by Drosophila homeoproteins: The role of DNA binding in functional activity and specificity. Development 124:4425-4433.

BLUMENTHAL, T. 1998. Gene clusters and polycistronic transcription in eukaryotes. BioEssays 20:480-487.

BONIFER, C. 2000. Developmental regulation of eukaryotic gene loci: Which cisregulatory information is required? Trends Genet 16:310-315.

BRAKEFIELD, P. M., J. GATES, D. KEYS, F. KESBEKE, P. J. WIJNGAARDEN, A. MONTEIRO, V. FRENCH, and S. B. CARROLL. 1996. Development, plasticity and evolution of butterfly eyespot patterns. Nature 384:236-242.

BREM, R. B., G. YVERT, R. CLINTON, and L. KRUGLYAK. 2002. Genetic dissection of transcriptional regulation in budding yeast. Science 296:752-755.

BRICKMAN, J. M., M. CLEMENTS, R. TYRELL, D. MCNAY, K. WOODS, J. WARNER, A. STEWART, R. S. P. BEDDINGTON, and M. DATTANI. 2001. Molecular effects of novel mutations in Hesx1/HESX1 associated with human pituitary disorders. Development 128:5189-5199.

BRITTEN, R. J. 1997. Mobile elements inserted in the distant past have taken on important functions. Gene 205:177-182.

BRITTEN, R. J., and E. H. DAVIDSON. 1969. Gene regulation for higher cells: A theory. Science 165:349-357.

BRITTEN, R. J., and E. H. DAVIDSON. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. Q Rev Biol 46:111-138.

BROSIUS, J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238:115-134.

BROWN, C. T., A. G. RUST, P. J. CLARKE et al. (and 8 co-authors). 2002. New

91

computational approaches for analysis of cis-regulatory neworks. Dev Biol 246:86-102.

BRUNETTI, C. R., J. E. SELEGUE, A. MONTEIRO, V. FRENCH, P. M. BRAKEFIELD, and S. B.

CARROLL. 2001. The generation and diversification of butterfly eyespot color patterns. Current Biol 11:1578-1585.

BUCKWOLD, V. E., Z. C. XU, T. S. B. YEN, and J. H. OU. 1997. Effects of a frequent doublenucleotide

basal core promoter mutation and its putative single-nucleotide precursor mutations on hepatitis B virus gene expression and replication. J Gen Virol 78:2055-2065.

BUDD, G. E. 1999. Does evolution in body patterning genes drive morphological change- or vice versa? BioEssays 21:326-332.

BUGGS, C., N. NASRIN, A. MODE, P. TOLLET, H.-F. ZHAO, J.-Å. GUSTAFSSON, and M. ALEXANDER-BRIDGES. 1998. IRE-ABP (Insulin Response Element-A Binding Protein), an SRY-like protein, inhibits C/EBPa (CCAAT/Enhancer-Binding Protein a) - stimulated expression of the sex-specific cytochrome P450 2C12 gene. Mol Endocrinol 12:1294-1309.

BÜRGLIN, T. 1997. Analysis of TALE superclass homeobox genes (MEIS, PBC, KNOX, Iriquois, TGIF) reveals a novel domain conserved between plants and animals. Nuc Acids Res 25:4173-4180.

BURKE, A. C., C. E. NELSON, B. A. MORGAN, and C. TABIN. 1995. Hox genes and the evolution of vertebrate axial morphology. Development 121:333-346.

BURSTIN, J., D. DE VIENNE, P. DUBREUIL, and C. DAMERVAL. 1994. Molecular markers and protein quantities as genetic descriptors in maize. I. Genetic diversity among 21 inbred lines. Theor Appl Genet 89:943-950.

BUSH, R. M., and K. PAIGEN. 1992. Evolution of □-glucuronidase regulation in the genus Mus. Evolution 46:1-15.

CALHOUN, V. C., A. STATHOPOULOS, and M. LEVINE. 2002. Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex. Proc Natl Acad Sci USA 99:9243-9247.

CAREY, M., and S. T. SMALE. 2000. Transcriptional regulation in eukaryotes: Concepts, strategies, and techniques. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

CARRIÓN, A. M., W. A. LINK, F. LEDO, B. MELLSTRÖM, and J. R. NARANJO. 1999. DREAM is a Ca2+-regulated transcriptional repressor. Nature 398:80-84.

CARROLL, S. B. 2000. Endless forms: The evolution of gene regulation and morphological diversity. Cell 101:577-580.

CARROLL, S. B. 1995. Homeotic genes and the evolution of arthropods and chordates. Nature 376:479-485.

CARROLL, S. B., J. K. GRENIER, and S. D. WEATHERBEE. 2001. From DNA to diversity: Molecular genetics and the evolution of animal design. Blackwell Science, Inc., Malden, MA.

CASPI, A., J. MCCLAY, T. E. MOFFITT, J. MILL, J. MARTIN, I. W. CRAIG, A. TAYLOR, and R. POULTON. 2002. Role of genotype in the cycle of violence in maltreated children. Science 297:851-854.

CAVALIERI, D., J. P. TOWNSEND, and D. L. HARTL. 2000. Manifold anomalies in gene expression in a vineyard isolate of Saccharomyces cerevisiae revealed by DNA microarray analysis. Proc Natl Acad Sci USA 97:12369-12374.

CAVENER, D. R. 1992. Transgenic animal studies on the evolution of genetic regulatory circuitries. BioEssays 14:237-244.

CEREB, N., and S. Y. YANG. 1994. The regulatory complex of HLA class I promoters exhibits locus-specific conservation with limited allelic variation. J Immunol 152:3873-3883.

CHEN, G., and A. J. COUREY. 2000. Groucho/TLE family proteins and transcriptional repression. Gene 249:1-16.

CHIU, C. H., H. SCHNEIDER, J. L. SLIGHTOM, D. L. GUMUCIO, and M. GOODMAN. 1997. Dynamics of regulatory evolution in primate beta-globin gene clusters: cis-mediated acquisition of simian gamma fetal expression patterns. Gene 205:47-57.

CHOO, Y., and A. KLUG. 1997. Physical basis of a protein-DNA recognition code. Curr Op Struct Biol 7:117-125.

CHRISTOPHIDES, G. K., I. LIVADRAS, C. SAVAKIS, and K. KOMITOPOLOU. 2000. Two medfly promoters that have originated by recent gene duplications drive distinct sex, tissue and temporal expression patterns. Genetics 156:173-182.

CHUNG, Y. D., H. C. KWON, K. W. CHUNG, S. J. KIM, K. J. KIM, and C. C. LEE. 1996. Identification of ovarian enhancer-binding factors which bind to ovarian enhancer 1 of the Drosophila genes yp1 and yp2. Mol Gen Genet 251:347-351.

CLARK, A. G. 1990. Genetic components of variation in energy storage in Drosophila melanogaster. Evolution 44:637-650.

COLLER, H. A., C. GRANDORI, P. TAMAYO, T. COLBERT, E. S. LANDER, R. N. EISENMAN, and T. R. GOLUB. 2000. Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion.

Proc Natl Acad Sci USA 97:3260-3265.

CONLON, F. L., L. FAIRCLOUGH, B. M. J. PRICE, E. S. CASEY, and J. C. SMITH. 2001. Determinants of T box protein specificity. Development 128:3749-3758.

COOPER, D. N. 1999. Human gene evolution. Academic Press, Inc., San Diego.

CORE, N., B. CHARROUX, A. MCCORMICK, C. VOLA, L. FASANO, M. P. SCOTT, and S. KERRIDGE. 1997. Transcriptional regulation of the Drososphila homeotic gene teashirt by the homeodomain protein Fushi tarazu. Mech Dev 68:157-172.

COSTA, P., and C. PLOMION. 1999. Genetic analysis of needle proteins in maritime pine. 2. Variation in protein accumulation. Silvae Genet. 48:146-150.

COUREY, A. J. 2001. Regulatory transcription factors and cis-regulatory regions. Pp. 17-34 in J. Locker, ed. Transcription Factors. Academic Press, Inc., San Diego.

COWELL, L. G., T. B. KEPLER, M. JANITZ, R. LAUSTER, and N. A. MITCHISON. 1998. The distribution of variation in regulatory gene segments, as present in MHC class II promoters. Genome Res 8:124-134.

COWEN, L. E., D. SANGLARD, D. CALABRESE, C. SIRJUSINGH, J. B. ANDERSON, and L. M. KOHN. 2000. Evolution of drug resistance in experimental populations of Candida albicans. J Bacteriol 182:1515-1522.

COWLES, C. R., J. N. HIRSHHORN, D. ALTSHULER, and E. S. LANDER. 2002. Detection of regulatory variation in mouse genes. Nat Genet 32:432-437.

CRAWFORD, D. L., J. A. SEGAL, and J. L. BARNETT. 1999. Evolutionary analysis of TATAless proximal promoter function. Mol Biol Evol 16:194-207.

CZERNY, T., G. SCHAFFNER, and M. BUSSLINGER. 1993. DNA sequence recognition by Pax proteins: bipartite structure of the paired domain and its binding site. Genes Dev

7:2048-2061.

D'ELIA, A. V., G. TELL, I. PARON, L. PELLIZZARI, R. LONIGRO, and G. DAMANTE. 2002. Missense mutations of human homeoboxes: A review. Human Mutation 18:361-374.

DABORN, P. J., J. L. YEN, M. R. BOGWITZ, G. L. GOFF, E. FEIL, S. JEFFERS, N. TIJET, T. PERRY,

D. HECKEL, P. BATTERHAM, R. FEYEREISEN, T. G. WILSON, and R. H. FFRENCHCONSTANT.

2002. A single P450 allele associated with insecticide resistance in

Drosophila. Science 297:2253-2225.

DAILEY, L., and C. BASILICO. 2001. Coevolution of HMG domains and homeodomains and the generation of transcriptional regulation by Sox/POU complexes. J Cell Physiol 186:315-328.

DAMERVAL, C., A. MAURICE, J. M. JOSSE, and D. DE VIENNE. 1994. Quantitative trait loci

underlying gene product variation: A novel perspective for analyzing regulation of genome expression. Genetics 137:289-301.

DARWIN, C. 1859. On the origin of species by means of natural selection. John Murray, London.

DAVIDSON, E. H. 2001. Genomic regulatory systems: Development and evolution. Academic Press, San Diego.

DAWES, R., I. DAWSON, F. FALCIANI, G. TEAR, and M. AKAM. 1994. Dax, a locust Hox gene

related to fushi-tarazu but showing no pair-rule expression. Development 120:1561-1572.

DAWSON, S. J., P. J. MORRIS, and D. S. LATCHMAN. 1996. A single amino acid change converts a repressor into an activator. J Biol Chem 271:11631-11633.

DE VIENNE, D., B. BOST, J. FIEVET, M. ZIVY, and C. DILLMANN. 2001. Genetic variability of
proteome expression and metabolic control. Plant Physiol Biochem 39:271-283.

DENIZOT, Y., E. PINAUD, C. AUPETIT, C. LE MORVAN, E. MAGNOUX, J. C. ALDIGIER, and M.
COGNE. 2001. Polymorphism of the human alpha1 immunoglobulin gene 3' enhancer hs1,2 and its relation to gene expression. Immunology 103:35-40.

DERMITZAKIS, E. T., and A. G. CLARK. 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: Conservation and turnover. Mol Biol Evol 19:1114-1121.

DEROBERTIS, E. M., and Y. SASAI. 1996. A common plan for dorsoventral patterning in Bilateria. Nature 380:37-40.

De Vienne et al. (1988) Genetic aspects of variation of protein amounts in maize and pea. Electrophoresis 9:742-750

DI GREGORIO, A., J. C. CORBO, and M. LEVINE. 2001. The regulation of forkhead/HNf-3 beta
expression in the Ciona embryo. Dev Biol 229:31-43.

DICKINSON, W. J. 1988. On the architecture of regulatory systems: Evolutionary insights and implications. BioEssays 8:204-208.

DILEONE, R. J., L. B. RUSSELL, and D. M. KINGSLEY. 1998. An extensive 3' regulatory region controls expression of Bmp5 in specific anatomical structures of the mouse embyro. Genetics 148:401-408.

DILLON, N., and P. SABBATTINI. 2000. Functional gene expression domains: Defining the functional unit of eukaryotic gene regulation. BioEssays 22:657-665.

DOBZHANSKY, T. 1936. Studies on hybrid sterility. II. Localization of sterility factors in Drosophila pseudoobscura hybrids. Genetics 21:113-135.

DOEBLEY, J., and L. LUKENS. 1998. Transcriptional regulators and the evolution of plant form. Plant Cell 10:1075-1082.

DROGE, P., and B. MULLER-HILL. 2001. High local protein concentrations at promoters:

strategies in prokaryotic and eukaryotic cells. BioEssays 23:179-183.

DUBOULE, D. 1994. Guidebook to the homeobox genes. Oxford University Press, Oxford.

DUBOULE, D., and A. S. WILKINS. 1998. The evolution of 'bricolage'. Trends Genetics 14:54-59.

DUDAREVA, N., L. CSEKE, V. M. BLANC, and E. PICHERSKY. 1996. Evolution of floral scent in Clarkia: Novel patterns of S-linalool synthase gene expression in the C. breweri flower. Plant Cell 8:1137-1148.

DYNAN, W. S. 1989. Modularity in promoters and enhancers. Cell 58:1-4.

ENARD, W., P. KHAITOVICH, J. KLOSE et al. (and 10 co-authors). 2002a. Intra- and interspecific variation in primate gene expression patterns. Science 296:340-343.

ENARD, W., M. PRZEWORSKI, S. E. FISHER, C. S. L. LAI, V. WIEBE, T. KITANO, A. P. MONACO, and S. PÄÄBO. 2002b. Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418:869-872.

FAIRALL, L., and J. W. R. SCHWABE. 2001. DNA binding by transcription factors. Pp. 65-84 in J. Locker, ed. Transcription Factors. Academic Press, Inc., San Diego.

FALCIANI, F., B. HAUSDORF, R. SCHRÖDER, M. AKAM, D. TAUTZ, R. DENELL, and S. BROWN. 1996. Class 3 Hox genes in insects and the origin of zen. Proc Natl Acad Sci USA 93:8479-8484.

FANG, H., and B. P. BRANDHORST. 1996. Expression of the actin gene family in embryos of the sea urchin Lytechinus pictus. Dev Biol 173:306-317.

FANG, S., A. TAKAHASHI, and C.-I. WU. 2002. A mutation in the promoter of desaturase 2 is correlated with sexual isolation between Drosophila behavioral races. Genetics 162:781-784.

FEREA, T. L., D. BOTSTEIN, P. O. BROWN, and R. F. ROSENZWEIG. 1999. Systematic changes in gene expression patterns following adaptive evolution in yeast. Proc Natl Acad Sci USA 96:9721-9726.

FERKOWICZ, M. J., and R. RAFF. 2001. Wnt gene expression in sea urchin develoment: Heterochronies associated with the evolution of developmental mode. Evol Dev 3:24-33.

FERRIGNO, O., T. VIROLLE, Z. DJABARI, J. P. ORTONNE, R. J. WHITE, and D. ABERDAM. 2001. Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. Nat Genet 28:77-81.

FERRIS, S. D., and G. S. WHITT. 1979. Evolution of the differential regulation of duplicate genes following polyploidization. J Mol Evol 12:267-317.

FISHER, R. A. 1930. The genetical theory of natural selection. Clarendon Press, Oxford.

FITCH, D. H. A. 1997. Evolution of male tail development in rhabditid nematodes related to Caenorhabditis elegans. Sys Biol 46:145-179.

FITCH, W. M., and E. MARKOWITZ. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet 4:579-593.

FLORES-SAAIB, R. D., S. JIA, and A. J. COUREY. 2001. Activation and repression by the Cterminal domain of Dorsal. Development 128:1869-1879.

FORCE, A., M. LYNCH, F. B. PICKETT, A. AMORES, Y.-L. YAN, and J. POSTLETHWAIT. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531-1545.

FOULKES, N. S., and P. SASSONE-CORSI. 1992. More is better: activators and repressors from the same gene. Cell 68:411-414.

FRASCH, M., X. CHEN, and T. LUFKIN. 1995. Evolutionary-conserved enhancers direct region-specific expression of the murine Hoxa-1 and Hoxa-2 loci in both mice and Drosophila. Development 121:957-974.

FRAZER, K. A., J. B. SHEEHAN, R. P. STOKOWSKI, X. CHEN, R. HOSSEINI, J. F. CHENG, S. P. FODOR, D. R. COX, and N. PATIL. 2001. Evolutionarily conserved sequences on human chromosome 21. Genome Res 11:1651-1659.

FRY, C. J., and P. J. FARNHAM. 1999. Context-dependent transcriptional regulation. J Biol Chem 274:29583-29586.

FU, Y.-X., and W.-H. LI. 1993. Statistical tests of neutrality of mutations. Genetics 133:693-709.

FULLERTON, S. M., A. BARTOSZEWICZ, G. YBAZETA, Y. HORIKAWA, G. I. BELL, K. K. KIDD, N. J. COX, R. R. HUDSON, and A. DI RIENZO. 2002. Geographic and haplotype structure of candidate type 2 diabetes-susceptibility variants at the calpain-10 locus. Amer J Hum Gen 70:1096-1106.

GALANT, R., and S. B. CARROLL. 2002. Evolution of a transcriptional repression domain in an insect Hox protein. Nature 415:910-913.

GEHRING, W. J., and K. IKEO. 1999. Pax6: Mastering eye morphogenesis and eye evolution. Trends Gen 15:371-377.

GERARD, M., J. ZAKANY, and D. DUBOULE. 1997. Interspecies exchange of a Hoxd enhancer in vivo induces premature transcription and anterior shift of the sacrum. Dev Biol 190:32-40.

GERBER, S., F. FABRE, and C. PLANCHON. 2000. Genetics of seed quality in soybean analysed by capillary gel electrophoresis. Plant Sci 152:181-189.

GERHART, J., and M. KIRSCHNER. 1997. Cells, embryos, and evolution: Toward a cellular and developmental understanding of phenotypic variation and evolutionary adaptability. Blackwell Science, Inc., Malden, MA.

GIBSON, G. 1996. Epistasis and pleiotropy as natural properties of transcriptional regulation. Theor Pop Biol 49:58-89.

GILBERT, S. F. 2001. Ecological developmental biology: Developmental biology meets the real world. Dev Biol 233:1-12.

GILBERT, S. F. 2000. Developmental biology. Sinauer Associates, Sunderland, MA.

GILL, G., and M. PTASHNE. 1987. Mutants of GAL4 protein altered in an activation function. Cell 51:121-126.

GILLESPIE, J. H. 1991. The causes of molecular evolution. Oxford University Press, New York.

GIORDANO, M., C. MARCHETTI, E. CHIORBOLI, G. BONA, and P. MOMIGLIANO RICHIARDI.

1997. Evidence for gene conversion in the generation of extensive polymorphism

in the promoter of the growth hormone gene. Hum Genet 100:249-255.

GLASS, C. K., D. W. ROSE, and M. G. ROSENFELD. 1997. Nuclear receptor coactivators. Curr Opin Cell Biol 9:222-232.

GONZALEZ, P., P. V. RAO, S. B. NUNEZ, and J. S. ZIGLER, JR. 1995. Evidence for independent recruitment of zeta-crystallin/quinone reductase (CRYZ) as a crystallin in camelids and hystricomorph rodents. Mol Biol Evol 12:773-781.

GOODYER, C. G., G. ZOGOPOLOS, G. SCHWARTZBAUER, H. ZHENG, G. N. HENDY, and R. K.

MENON. 2001. Organization and evolution of the human growth hormone receptor 5'-flanking region. Endocrinology 142:1923-1934.

GRANDORI, C., S. M. COWLEY, L. P. JAMES, and R. N. EISENMAN. 2000. The Myc/Max/Mad

network and the transcriptional control of cell behavior. Annu Rev Cell Dev Biol 16:653-699.

GRAY, S., and M. LEVINE. 1996. Transcriptional repression in development. Curr Op Cell Biol 8:358-364.

GRBIC, M., L. M. NAGY, and M. R. STRAND. 1998. Polyembryonic insect development: Insect pattern formation in a cellularised environment. Dev Genes Evol 208:69-81.

GROSVELD, F., M. ANTONIOU, M. BERRY et al. (and 13 co-authors). 1993. The regulation of human globin gene switching. Phil Trans R Soc Lond B 339:183-191.

GSTAIGER, M., L. KNOEPFL, O. GEORGIEV, W. SCHAFFNER, and C. M. HOVENS. 1995. A B-cell co-activator of octamer-binding transcription factors. Nature 373:360-362.

GU, Z., D. NICOLAE, H. H.-S. LU, and W.-H. LI. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet 18:609-613.

GUARDIOLA, J., A. MAFFEI, R. LAUSTER, N. A. MITCHISON, R. S. ACCOLLA, and S. SARTORIS. 1996. Functional significance of polymorphism among MHC class II gene promoters. Tissue Antigens 48:615-625.

HAHN, M. W., M. D. RAUSHER, and C. W. CUNNINGHAM. 2002. Distinguishing between selection and population expansion in an experimental lineage of bacteriophage T7. Genetics 161:11-20.

HAHN, M. W., J. E. STAJICH, and G. A. WRAY. 2003. The effects of selection against spurious transcription factor binding sites. Mol Biol Evol, in press.

HAMBLIN, M. T., and A. DI RIENZO. 2000. Detection of the signature of natural selection in humans: Evidence from the Duffy blood group locus. Amer J Hum Gen 66:1669-1679.

HANCOCK, J., P. SHAW, F. BENNETON, and G. DOVER. 1999. High sequence turnover in the regulatory regions of the developmental gene hunchback in insects. Mol Biol Evol 16:253-265.

HANES, S. D., G. RIDDIHOUGH, D. ISH-HOROWICZ, and R. BRENT. 1994. Specific DNA recognition and intersite spacing are critical for action of the bicoid morphogen.

Mol Cell Biol 14:3364-3375.

HANNA-ROSE, W., and U. HANSEN. 1996. Active repression mechanisms of eukaryotic transcription factors. Trends Gen 12:229-234.

HARIRI, A. R., V. S. MATTAY, A. TESSITORE, B. KOLACHANA, F. FERA, D. GOLDMAN, M. F. EGAN, and D. R. WEINBERGER. 2002. Serotonin transporter genetic variation and the response of the human amygdala. Science 297:400-403.

HARRISON, S. C. 1991. A structural taxonomy of DNA-binding domains. Nature 353:715-719.

HAUDEK, S. B., B. E. NATMESSNIG, H. REDL, G. SCHLAG, and B. P. GIROIR. 1998. Genetic sequences and transcriptional regulation of the TNFA promoter: Comparison of human and baboon. Immunogenetics 48:202-207.

HILL, T. A., C. D. DAY, S. C. ZONDLO, A. G. THACKERAY, and V. F. IRISH. 1998. Discrete spatial and temporal cis-acting elements regulate transcription of the Arabidopsis floral homeotic gene APETALA3. Development 125:1711-1721.

HIZVER, J., H. ROZENBERG, F. FROLOW, D. RABINOVICH, and Z. SHAKKED. 2001. DNA bending by an adenine-thymine tract and its role in gene regulation. Proc Natl Acad Sci USA 98:8490-8495.

HODGKIN, J. 1992. Genetic sex determination mechanisms and evolution. BioEssays 14:253-261.

HOLLAND, N. D., and L. Z. HOLLAND. 1999. Amphioxus and the utility of molecular genetic data for hypothesizing body part homologies between distantly related animals. Amer Zool 39:630-640.

HOLSTEGE, F. C. P., E. G. JENNINGS, J. J. WYRICK, T. I. LEE, C. J. HENGARTNER, M. R. GREEN, T. R. GOLUB, E. S. LANDER, and R. A. YOUNG. 1998. Dissecting the regulatory circuitry

of a eukaryotic genome. Cell 95:717-728.

HOPE, I. A., and K. STRUHL. 1986. Functional dissection of a eukaryotic transcriptional activator, GCN4 of yeast. Cell 46:885-894.

HOUCHENS, C. R., W. MONTIGNY, L. ZELTSER, L. DAILEY, J. M. GILBERT, and N. H. HEINTZ.

2000. The dhfr ori beta-binding protein RIP60 contains 15 zinc fingers: DNA binding and looping by the central three fingers and an associated proline-rich region. Nuc Acids Res 28:570-581.

HUDSON, R. R., M. KREITMAN, and M. AGUADE. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153-159.

HUNT, G. M., D. JOHNSON, and C. T. TIEMESSE. 2001. Characterisation of the long terminal

repeat regions of South African human immunodeficiency virus type 1 isolates. Virus Genes 23:27-34.

INDOVINA, P., F. MEGIORNI, P. FERRANTE, I. APOLLONIO, F. PETRONZELLI, and M. C. MAZZILLI. 1998. Different binding of NF-Y transcriptional factor to DQA1 promoter variants. Hum Immunol 59:758-767.

IYER, V. R., C. E. HORAK, C. S. SCAFE, D. BOTSTEIN, M. SNYDER, and P. O. BROWN. 2001.

Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 409:533-538.

JACKSON, S. P., and R. TJIAN. 1988. O-glycosylation of eukaryotic transcription factors: implications for mechanisms of transcriptional regulation. Cell 55:125-133.

JACKSON-FISHER, A. J., C. CHITIKILA, M. MITRA, and B. F. PUGH. 1999. A role for TBP dimerization in preventing unregulated gene expression. Mol Cell 3:717-727.

JACOB, F. 1977. Evolution and tinkering. Science 196:1161-1166.

JACOB, F., and J. MONOD. 1961. On the regulation of gene activity. Cold Spring Harbor Symp Quant Biol 26:193-211.

JACOBS, J. J., and M. VAN LOHUIZEN. 1999. Cellular memory of transcriptional states by Polycomb-group proteins. Sem Cell Dev Biol 10:227-235.

JAMES, L., and R. N. EISENMAN. 2002. Myc and Mad bHLHZ domains posess identical DNA-binding specificies but only partially overlapping functions in vivo. Proc Natl Acad Sci USA 99:10429-10434.

JAREBORG, N., E. BIRNEY, and R. DURBIN. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res 9:815-824.

JAYNES, J. B., and P. H. O'FARRELL. 1991. Active repression of transcription by the engrailed homeodomain protein. EMBO J 10:1427-1433.

JEENINGA, R. E., M. HOOGENKAMP, M. ARMAND-UGON, M. DE BAAR, K. VERHOEF, and B. BERKHOUT. 2000. Functional differences between the long terminal repeat transcriptional promoters of human immunodeficiency virus type 1 subtypes A through G. J Virol 74:3740-3751.

JENKINS, D. L., C. A. ORTORI, and J. F. Y. BROOKFIELD. 1995. A test for adaptive change in DNA sequences controlling transcription. Proc R Soc Lond B Biol Sci 261:203-207.

JIN, W., R. M. RILEY, R. D. WOLFINGER, K. P. WHITE, G. PASSADOR-GURGEL, and G. GIBSON. 2001. The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. Nat Gen 29:389-395.

JOHNSON, N. A., and A. H. PORTER. 2000. Rapid speciation via parallel, directional selection on regulatory genetic pathways. J Theor Biol 205:527-542.

JONES, P. A., and D. TAKAI. 2001. The role of DNA methylation in mammalian epigenetics. Science 293:1068-1070.

JONES, S., P. VAN HEYNINGEN, H. M. BERMAN, and J. M. THORNTON. 1999. Protein-DNA interactions: a structural analysis. J Mol Biol 287:877-896.

JORDAN, I. K., and J. F. MCDONALD. 1998. Interelement selection in the regulator region of the copia retrotransposon. J Mol Evol 47:670-676.

KADOSH, D., and K. STRUHL. 1998. Targeted recruitment of the Sin3-Rpd3 histone deacetylase complex generates a highly localized domain of repressed chromatin in vivo. Mol Cell Biol 18:5121-5127.

KAJIYA, Y., K. HAMASAKI, K. NAKATA et al. (and 8 co-authors). 2001. A long-term followup analysis of serial core promoter and precore sequences in Japanese patients chronically infected by hepatitis B virus. Digest Diseases Sci 46:509-515.

KAMMANDEL, B., K. CHOWDHURY, A. STOYKOVA, S. APARICIO, S. BRENNER, and P. GRUSS. 1999. Distinct cis-essential modules direct the time-space pattern of the Pax6 gene activity. Dev Biol 205:79-97.

KARP, C. L., A. GRUPE, E. SCHADT et al. (and 10 co-authors). 2000. Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. Nat Immunol 1:221-226.

KAYO, T., D. B. ALLISON, R. WEINDRUCH, and T. A. PROLLA. 2001. Influences of aging and caloric restriction on the transcriptional profile of skeletal muscle from rhesus monkeys. Proc Natl Acad Sci USA 98:5093-5098.

KAZAZIAN, H. H. 1990. The thalassemia syndromes: Molecular basis and prenatal diagnosis in 1990. Sem Hematol 27:209-228.

KEYS, D. N., D. L. LEWIS, J. E. SELEGUE, B. J. PEARSON, L. V. GOODRICH, R. L. JOHNSON, J. GATES, M. P. SCOTT, and S. B. CARROLL. 1999. Recruitment of a hedgehog regulatory circuit in butterfly eyespot evolution. Science 283:532-534.

KIDWELL, M. G., and D. LISCH. 1997. Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci USA 94:7704-7711.

KIM, J. 2001. Macro-evolution of the hairy enhancer in Drosophila species. J Exp Zool 291:175-185.

KIM, J., J. Q. KERR, and G.-S. MIN. 2000. Molecular heterochrony in the early development of Drosophila. Proc Natl Acad Sci USA 97:212-216.

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

KING, M. C., and A. C. WILSON. 1975. Evolution at two levels in humans and chimpanzees. Science 188:107-116.

KIRCHHAMER, C. V., L. D. BOGARAD, and E. H. DAVIDSON. 1996. Developmental expression of synthetic cis-regulatory systems composed of spatial control elements from two different genes. Proc Natl Acad Sci USA 93:13849-13854.

KIRCHHAMER, C. V., C.-H. YUH, and E. H. DAVIDSON. 1996. Modular cis-regulatory organization of developmentally expressed genes: two genes transcribed territorially in the sea urchin embryo, and additional examples. Proc Natl Acad Sci USA 93:9322-9328.

KISSINGER, J. C., and R. A. RAFF. 1998. Evolutionary changes in sites and timing of actin gene expression in embryos of the direct- and indirect-developing sea urchins, Heliocidaris erythrogramma and H. tuberculata. Dev Genes Evol 208:82-93.

KLARENBERG, A. J., K. SIKKEMA, and W. SCHARLOO. 1987. Functional significance of regulatory map and structural amy variants in Drosophila melanogaster. Heredity 58:383-389.

KLEIN, W. H., and X. T. LEE. 1999. Function and evolution of Otx proteins. Biochem Biophys Res Comm 258:229-233.

KLOSE, J., C. NOCK, M. HERRMANN et al. (and 10 co-authors). 2002. Genetic analysis of the mouse brain proteome. Nat Genet 30:385-393.

KMITA, M., T. KONDO, and D. DUBOULE. 2000. Targeted inversion of a polar silencer within the HoxD complex re-allocates domains of enhancer sharing. Nat Genet

26:451-454.

Knight et al. (1999) X-linked dyskeratosis congenita is predominantly caused by missense mutations in the DKC1 gene. Am. J. Hum. Genet. 65:50-58

KNOEPFLER, P. S., and M. P. KAMPS. 1995. The pentapeptide motif of hox proteins is required for cooperative DNA binding with pbx1, physically contacts pbx1 and enhances DNA binding by pbx1. Mol Cell Biol 15:5811-5819.

KOPP, A., I. DUNCAN, and S. B. CARROLL. 2000. Genetic control and evolution of sexually dimorphic characters in Drosophila. Nature 408:553-559.

KORNEEV, S., and M. O'SHEA. 2002. Evolution of nitric oxide synthase regulatory genes by DNA inversion. Mol Biol Evol 19:1228-1233.

KRAMER, S. G., T. M. JINKS, P. SCHEDL, and J. P. GERGEN. 1999. Direct activation of Sexlethal

transcription by the Drosophila runt protein. Development 126:191-200.

KURAS, L., and K. STRUHL. 1999. Binding of TBP to promoters in vivo is stimulated by activators and requires Pol II holoenzyme. Nature 399:609-613.

LANDER, E. S., L. M. LINTON, B. BIRREN et al. (and 239 co-authors). 2001. Initial sequencing

and analysis of the human genome. Nature 409:860-921.

LATCHMAN, D. S. 1998. Eukaryotic transcription factors. Academic Press, San Diego.

LAURIE-AHLBERG, C. C., and G. C. BEWLEY. 1983. Naturally occurring genetic variation affecting the expression of sn-glycerol-3-phosphate dehydrogenase in Drosophila melanogaster. Biochem Genet 21:943-961.

LAURIE-AHLBERG, C. C., G. MARONI, G. C. BEWLEY, J. C. LUCCHESI, and B. S. WEIR. 1980.

Quantitative genetic variation of enzyme activities in natural populations of Drosophila melanogaster. Proc Natl Acad Sci USA 77:1073-1077.

LAWTON-RAUH, A. L., E. R. ALVAREZ-BUYLLA, and M. D. PURUGGANAN. 2000. Molecular

evolution of flower development. Trends Ecol Evol 15:144-149.

LEE, H., S. N. CHO, H. E. BANG, J. H. LEE, G. H. BAI, S. J. KIM, and J. D. KIM. 2000. Exclusive mutations related to isoniazid and ethionamide resistance among Mycobacterium tuberculosis isolates from Korea. Eur J Clin Microbiol Infect Dis 17:508-511.

LEE, J. E. 1997. Basic helix-loop-helix genes in neural development. Curr Opin Neurobiol 7:13-20.

LEE, T. I., N. J. RINALDI, R. ROBERT et al. (and 18 co-authors). 2002. Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298:799-804.

LEE, T. I., and R. A. YOUNG. 2000. Transcription of eukaryotic protein-coding genes. Annu Rev Genet 34:77-137.

LEMON, B., and R. TJIAN. 2000. Orchestrated response: A symphony of transcription factors for gene control. Genes Dev 14:2551-2569.

LERMAN, D. N., P. MICHALAK, A. B. HELIN, B. R. BETTENCOURT, and M. E. FEDER. 2003.

Modification of heat-shock gene expression in Drosophila melanogaster populations via transposable elements. Mol Biol Evol 20:135-144.

LETTICE, L. A., T. HORIKOSHI, S. J. H. HEANEY et al. (and 18 co-authors). 2002. Disruption

of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc Natl Acad Sci U S A 99:7548-7553.

LEWIN, B. 2000. Genes VII. Oxford University Press, Oxford.

LEWIS, E. B. 1978. Gene complex controlling segmentation in Drosophila. Nature 276:565-570.

LI, Q. M., and S. A. JOHNSTON. 2001. Are all DNA binding and transcription regulation by an activator physiologically relevant? Mol Cell Biol 21:2467-2474.

LI, W.-H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

LI, W. W., M. M. DAMMERMAN, J. D. SMITH, S. METZGER, J. L. BRESLOW, and T. LEFF. 1995.

Common genetic variation in the promoter of the human apo CIII gene abolishes regulation by insulin and may contribute to hypertriglyceridemia. J Clin Invest 96:2601-2605.

LI, X., and M. NOLL. 1994. Evolution of distinct developmental functions of three Drosophila genes by acquisition of different cis-regulatory regions. Nature 367:83-87.

LI, Y., J. P. BERNOT, C. ILLINGWORTH et al. (and 7 co-authors). 2001. Gene conversion within regulatory sequences generates maize r alleles with altered gene expression. Genetics 159:1727-1740.

LIANG, Z., and M. D. BIGGIN. 1998. Eve and ftz regulate a wide array of genes in blastoderm embryos: the selector homeoproteins directly or indirectly regulate most genes in Drosophila. Development 125:4471-4482.

LIEB, J. D., X. LIU, D. BOTSTEIN, and P. O. BROWN. 2001. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat Genet 28:327-334.

LIU, T., J. WU, and F. HE. 2000. Evolution of cis-acting elements in 5' flanking regions of vertebrate actin genes. J Mol Evol 50:22-30.

LOCKER, J. 2001. Transcription factors. Academic Press, Inc., San Diego.

LONG, M., W. WANG, and J. ZHANG. 1999. Origin of new genes and source for N-terminal domain of the chimerical gene, jingwei, in Drosophila. Gene 238:135-141.

LOOTS, G. G., R. M. LOCKSLEY, C. M. BLANKESPOOR, Z. E. WANG, W. MILLER, E. M. RUBIN, and K. A. FRAZER. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science 288:136-140.

LOVE, J. J., X. LI, D. A. CASE, K. GEISE, R. GROSSCHEDL, and P. E. WRIGHT. 1995. Structural basis for DNA bending by the architectural transcription factor LEF-1. Nature

376:791-795.

LOWE, C. J., and G. A. WRAY. 1997. Radical alterations in the roles of homeobox genes during echinoderm evolution. Nature 389:718-721.

LUDWIG, M. Z. 2002. Functional evolution of noncoding DNA. Curr Op Genet Dev 12:634-639.

LUDWIG, M. Z., C. BERGMAN, N. H. PATEL, and M. KREITMAN. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403:564-567.

LUDWIG, M. Z., and M. KREITMAN. 1995. Evolutionary dynamics of the enhancer region of even-skipped in Drosophila. Mol Biol Evol 12:1002-1011.

LUDWIG, M. Z., N. H. PATEL, and M. KREITMAN. 1998. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: Rules governing conservation and change. Development 125:949-958.

LUFKIN, T. 2001. Developmental control by Hox transcriptional regulators and their cofactors. Pp. 215-235 in J. Locker, ed. Transcription Factors. Academic Press, Inc., San Diego.

LUTZ, B., H. C. LU, G. EICHELE, D. MILLER, and T. C. KAUFMAN. 1996. Rescue of Drosophila labial null mutant by the chicken ortholog Hoxb-1 demonstrates that the function of Hox genes is phylogenetically conserved. Genes Dev 10:176-184.

MA, X., D. YUAN, K. DIEPOLD, T. SCARBOROUGH, and J. MA. 1996. The Drosophila morphogenic protein Bicoid binds DNA cooperatively. Development 122:1195-1206.

MADURO, M., and D. PILGRIM. 1996. Conservation of function and expression of unc-119 from two Caenorhabditis species despite divergence of non-coding DNA. Gene 183:77-85.

MAHMOUDI, T., and C. P. VERRIJZER. 2001. Chromatin silencing and activation by Polycomb and trithorax group proteins. Oncogene 20:3055-3066.

MAI, X., S. CHOU, and K. STRUHL. 2000. Preferential accessibility of the yeast his3

promoter is determined by a general property of DNA sequence, not by specific elements. Mol Cell Biol 20:6668-6676.

MANZANARES, M., H. WADA, N. ITASAKI, P. A. TRAINOR, R. KRUMLAUF, and P. W. HOLLAND. 2000. Conservation and elaboration of Hox gene regulation during evolution of the vertebrate head. Nature 408:854-857.

MARGARIT, E., A. GUILLÉN, C. BEBORDOSA, J. VIDAL-TABOADA, M. SÁNCHEZ, F. BALLESTA, and R. OLIVA. 1998. Identification of conserved potentially regulatory sequences of the SRY gene from 10 different species of mammals. Biochem Biophys Res Comm 245:370-377.

MASTICK, G. S., R. MCKAY, T. OLIGINO, K. DONOVAN, and A. J. LÓPEZ. 1995. Identification of target genes regulated by homeotic proteins in Drosophila melanogaster through genetic selection of Ultrabithorax protein-binding sites in yeast. Genetics 139:349-363.

MATHIAS, J. R., H. L. ZHONG, H. H. JIN, and A. K. VERSHON. 2001. Altering the DNAbinding specificity of the yeast Mat alpha 2 homeodomain protein. J Biol Chem 276:32696-32703.

MATSUO, Y., and T. YAMAZAKI. 1984. Genetic analysis of natural populations of Drosophila melanogaster in Japan. IV. Natural selection on the inducibility, but not on the structural genes, of amylase loci. Genetics 108:879-896.

MCDONALD, J. H., and M. KREITMAN. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature:652-654.

MCKINNEY, M. L., and K. J. MCNAMARA. 1991. Heterochrony: the evolution of ontogeny. Plenum Press, New York.

METHERALL, J. E., F. P. GILLESPIE, and B. G. FORGET. 1988. Analyses of linked beta-globin genes suggest that nondeletion forms of hereditary persistence of fetal hemoglobin are bona fide switching mutants. Am J Hum Genet 42:476-481.

MEYER, C. G., J. MAY, A. J. LUTY, B. LELL, and P. G. KREMSNER. 2002. TNFa-308A associated with shorter invervals of Plasmodium falciparum reinfections. Tissue Antigens 59:287-292.

MILO, R., S. SHEN-ORR, S. ITZKOVITZ, D. KASHTAN, D. CHKLOVSKII, and U. ALON. 2002. Network motifs: simple building blocks of complex networks. Science 298:824-827.

MITSIALIS, S. A., and F. C. KAFATOS. 1985. Regulatory elements controlling chorion gene expression are conserved between flies and moths. Nature 317:453-456.

MIYASHITA, N. T. 2001. DNA variation in the 5' upstream region of the Adh locus of the wild plants Arabidopsis thaliana and Arabis gemmifera. Mol Biol Evol 18:164-171.

MODY, M., Y. CAO, Z. CUI et al. (and 7 co-authors). 2001. Genome-wide gene expression profiles of the developing mouse hippocampus. Proc Natl Acad Sci USA 98:8862-8867.

MONTANO, M. A., C. P. NIXON, T. NDUNG'U, H. BUSSMANN, V. A. NOVITSKY, D. DICKMAN, and M. ESSEX. 2000. Elevated tumor necrosis factor-alpha activation of human immunodeficiency virus type 1 subtype C in Southern Africa is associated with an NF-kappaB enhancer gain-of-function. J Infect Dis 181:76-81.

MONTANO, M. A., V. A. NOVITSKY, J. T. BLACKARD, N. L. CHO, D. A. KATZENSTEIN, and M. ESSEX. 1997. Divergent transcriptional regulation among expanding human immunodeficiency virus type 1 subtypes. J Virol 71:8657-8665.

MORISHIMA, A. 1998. Identification of preferred binding sites of a light-inducible DNAbinding factor (MNF1) within 5'-upstream sequence of C4-type phosphoenolpyruvate carboxylase gene in maize. Plant Mol Biol 38:633-646.

MOUCHEL-VIELH, E., M. BLIN, C. RIGOLOT, and J. S. DEUTSCH. 2002. Expression of a homologue of the fushi-tarazu (ftz) gene in a cirriped crustacean. Evol Dev 4:76-85.

MÜLLER, G. B., and G. P. WAGNER. 1991. Novelty in evolution: Restructuring the concept. Annu Rev Ecol Sys 22:229-256.

MULLER, H. J. 1942. Isolating mechanisms, evolution, and temperature. Biol Symposia 6:71-125.

NAGANAWA, S., H. N. GINSBERG, R. M. GLICKMAN, and G. S. GINSBURG. 1997. Intestinal transcription and synthesis of apolipoprotein AI is regulated by five natural polymorphisms upstream of the apolipoprotein CIII gene. J Clin Invest 99:1958-1965.

NAKAYAMA, E. E., L. MEYER, A. IWAMOTO, A. PERSOZ, Y. NAGAI, C. ROUZIOUX, J.-F. DELFRAISSY, P. DEBRE, D. MCILROY, I. THEODOROU, T. SHIODA, and S. S. GROUP. 2002. Protective effect of interluekin-4 -589T polymorphism on human immunodeficiency virus type 1 disease progression: relationship with virus load. J Inf Diseases 185:1183-1186.

NARLIKAR, G. J., H.-Y. FAN, and R. E. KINGSTON. 2002. Cooperation between complexes that regulate chromatin structure and transcription. Cell 108:475-487.

NAYLOR, L. H., and E. M. CLARK. 1990. d(TG)n.d(CA)n sequences upstream of the rat prolactin gene form Z-DNA and inhibit gene transcription. Nuc Acids Res 18:1595-1601.

NEZNANOV, N., A. UMEZAWA, and R. G. OSHIMA. 1997. A regulatory element within a coding exon modulates keratin 18 gene expression in transgenic mice. J Biol Chem

272:27549-27557.

NIELSEN, L. B., D. KAHN, T. DUELL, H. U. WEIER, S. TAYLOR, and S. G. YOUNG. 1998. Apolipoprotein B gene expression in a series of human apolipoprotein B transgenic mice generated with recA-assisted restriction endonuclease cleavage-modified bacterial artificial chromosomes. An intestine- specific enhancer element is located between 54 and 62 kilobases 5' to the structural gene. J Biol Chem 273:21800-21807.

NURMINSKY, D. I., E. N. MORIYAMA, E. R. LOZOVSKAYA, and D. L. HARTL. 1996. Molecular phylogeny and genome evolution in the Drosophila virilis species group: Duplications of the alcohol dehydrogenase gene. Mol Biol Evol 13:132-149.

NURMINSKY, D. I., M. V. NURMINSKAYA, D. DEAGUIAR, and D. L. HARTL. 1998. Selective sweep of a newly evolved sperm-specific gene in Drosophila. Nature 396:572-575.

ODGERS, W. A., M. J. HEALY, and J. G. OAKESHOTT. 1995. Nucleotide polymorphism in the 5' promoter region of esterase 6 in Drosophila melanogaster and its relationship to enzyme activity variation. Genetics 141:215-222.

OHLER, U., and H. NIEMANN. 2001. Identification and analysis of eukaryotic promoters: recent computational approaches. Trends in Genetics 17:56-60.

OHTA, T. 1992. The nearly neutral theory of molecular evolution. Annu Rev Ecol Sys 23:263-286.

OHTSUKI, S., M. LEVINE, and H. N. CAI. 1998. Different core promoters possess distinct regulatory activities in the Drosophila embryo. Genes Dev 12:547-556.

OLEKSIAK, M. F., G. A. CHURCHILL, and D. L. CRAWFORD. 2002. Variation in gene expression within and among natural populations. Nat Genet 32:261-266.

ONYANGO, P., W. MILLER, J. LEHOCZKY, C. T. LEUNG, B. BIRREN, S. WHEELAN, K. DEWAR, and A. P. FEINBERG. 2000. Sequence and comparative analysis of the mouse 1-

megabase region orthologous to the human 11p15 imprinted domain. Genome Res 10:1697-1710.

ORPHANIDES, G., T. LAGRANGE, and D. REINBERG. 1998. The general transcription factors of RNA polymerase II. Genes and Development 10:2657-2683.

ORR, H. A., and D. C. PRESGRAVES. 2000. Speciation by postzygotic isolation: forces, genes and molecules. BioEssays 22:1085-1094.

OSADA, S., H. YAMAMOTO, T. NISHIHARA, and M. IMAGAWA. 1996. DNA binding specificity of the CCAAT/enhancer-binding protein transcription factor family. J Biol Chem 271:3891-3896.

PAIGEN, K. 1989. Experimental approaches to the study of regulatory evolution. Am. Nat. 134:440-458.

PANGANIBAN, G., S. M. IRVINE, C. LOWE et al. (and 11 co-authors). 1997. The origin and evolution of animal appendages. Proc Natl Acad Sci USA 94:5162-5166.

PAPENBROCK, T., R. L. PETERSON, R. S. LEE, T. HSU, A. KUROIWA, and A. AWGULEWITSCH. 1998. Murine Hoxc-9 gene contains a structurally and functionally conserved enhancer. Dev Dyn 212:540-547.

PAQUETTE, J., N. GIANNOUKAKIS, C. POLYCHRONAKOS, P. VAFIADIS, and C. DEAL. 1998. The INS 5' variable number of tandem repeats is associated with IGF2 expression in humans. J Biol Chem 273:14158-14164.

PARKS, A. L., B. A. PARR, J.-E. CHIN, D. S. LEAF, and R. A. RAFF. 1988. Molecular analysis of heterochronic changes in the evolution of direct developing sea urchins. J Evol Biol 1:27-44.

PATRINOS, G. P., P. KOLLIA, A. LOUTRADI-ANAGNOSTOU, D. LOUKOPOULOS, and M. N.
PAPADAKIS. 1998. The Cretan type of non-deletional hereditary persistence of fetal hemoglobin [A gamma-158C-->T] results from two independent gene conversion events. Hum Genet 102:629-634.

PETRONZELLI, F., A. KIMURA, P. FERRANTE, and M. C. MAZZILLI. 1995. Polymorphism in
the upstream regulatory region of DQA1 gene in the Italian population. Tissue Antigens 45:258-263.

Pfister et al. (1982) Expression of beta-glucuronidase haplotypes in prototype and congenic mouse strains. Biochem Genet. 20:519-536

PIANO, F., M. J. PARISI, R. KARESS, and M. P. KAMBYSELLIS. 1999. Evidence for redundancy
but not trans factor-cis element coevolution in the regulation of Drosophila Yp genes. Genetics 152:605-616.

PINSONNEAULT, J., B. FLORENCE, H. VAESSIN, and W. MCGINNIS. 1997. A model for extradenticle function as a switch that changes HOX proteins from repressors to activators. EMBO J 16:2032-2042.

PIRKKALA, L., P. NYKANEN, and L. SISTONEN. 2001. Roles of the heat shock transcription
factors in regulation of the heat shock response and beyone. FASEB J 15:1118-1131.

PLAZA, S., S. SAULE, and C. DOZIER. 1999. High conservation of cis-regulatory elements between quail and human for the Pax-6 gene. Dev Genes Evol 209:165-173.

POWELL, J. R. 1979. Population genetics of Drosophila amylase. II. Geographic patterns in D. pseudoobscura. Genetics 92:613-622.

POWELL, J. R., and J. M. LICHTENFELS. 1979. Population genetics of Drosophila amylase. I.
Genetic control of tissue-specific expression in D. pseudoobscura. Genetics 92:603-612.

Prokunina et al. (2002) A regulatory polymorphism in PDCD1 is associated with susceptibility to systemic lupus erythematosus in humans. Nat. Genet. 32:666-669

PUGH, B. F. 2001. RNA polymerase II transcription machinery. Pp. 1-16 in J. Locker, ed. Transcription Factors. Academic Press, Inc., San Diego.

PURUGGANAN, M. D. 2000. The molecular population genetics of regulatory genes. Mol Ecol 9:1451-1461.

QUIRING, R., U. WALLDORF, U. KLOTER, and W. J. GEHRING. 1994. Homology of the eyeless gene of Drosophila to the small eye gene in mice and aniridia in humans. Science 265:785-789.

RAFF, R. A. 1996. The shape of life: Genes, development, and the evolution of animal form. The University of Chicago Press, Chicago.

RAFF, R. A., and T. C. KAUFMAN. 1983. Embryos, genes, and evolution: The developmental-genetic basis of evolutionary change. Macmillan Publishing Co., Inc., New York.

REGIER, J. C., and N. S. VLAHOS. 1988. Heterochrony and the introduction of novel modes of morphogenesis during the evolution of moth choriogenesis. J Mol Evol 28:19-31.

REINBERG, D., G. OPRHANIDES, R. EBRIGHT et al. (and 26 co-authors). 1998. The RNA polymerase II general transcription factors: past, present, and future. Cold Spring Harbor Symposium on Quatitative Biology 63:83-103.

REN, B., F. ROBERT, J. J. WYRICK et al. (and 11 co-authors). 2000. Genome-wide location and function of DNA binding proteins. Science 290:2306-2309.

RICHARDS, E. J., and S. C. R. ELGIN. 2002. Epigenetic codes for heterochromatin silencing: rounding up the usual suspects. Cell 108:489-500.

RIECHMANN, J. L., M. WANG, and E. M. MEYEROWITZ. 1996. DNA-binding properties of Arabidopsis MADS domain homeotic proteins APETALA1, APETALA3,

PISTILLATA and AGAMOUS. Nuc Acids Res 24:3134-3141.

RIFKIN, S. A., J. KIM, and K. P. WHITE. in press. Evolution of gene expression in the Drosophila melanogaster subgroup. Nat Genet.

RINDER, H., A. THOMSCHKE, S. RUSCH-GERDES, G. BRETZEL, K. FELDMANN, M. RIFAI, and T. LOSCHER. 1998. Significance of ahpC promoter mutations for the prediction of isoniazid resistance in Mycobacterium tuberculosis. Eur J Clinical Microbiol Infect Dis 17:508-511.

ROBERTS, S. B., N. SEGIL, and H. HEINTZ. 1991. Differential phosphorylation of the transcription factor Oct1 during the cell cycle. Science 253:1022-1026.

ROBIN, C., R. F. LYMAN, A. D. LONG, C. H. LANGLEY, and T. F. C. MACKAY. 2002. hairy: a quantitative trait locus for Drosophila sensory bristle number. Genetics 162:155-164.

ROCKMAN, M. V., and G. A. WRAY. 2002. Abundant raw material for cis-regulatory evolution in humans. Mol Biol Evol 19:1981-1990.

ROMANO, L. A., and G. A. WRAY. submitted. Conservation of Endo16 expression in sea urchins despite evolutionary divergence in both cis and trans-acting elements of transcriptional regulation.

ROMEY, M. C., C. GUITTARD, J. P. CHAZALETTE et al. (and 11 co-authors). 1999. Complex allele [-102T>A+S549R(T>G)] is associated with milder forms of cystic fibrosis than allele S549R[T>G] alone. Hum Genet 105:145-150.

ROMEY, M. C., N. PALLARES-RUIZ, A. MANGE, C. METTLING, R. PEYTAVI, J. DEMAILLE, and M. CLAUSTRES. 2000. A naturally occurring sequence variation that creates a YY1 element is associated with increased cystic fibrosis transmembrane conductance regulator gene expression. J Biol Chem 275:3561-3567.

RONSHAUGEN, M., N. MCGINNIS, and W. MCGINNIS. 2002. Hox protein mutation and

macroevolution of the insect body plan. Nature 415:914-917.

ROSS, J. L., P. P. FONG, and D. R. CAVENER. 1994. Correlated evolution of the cis-acting regulatory elements and developmental expression of the Drosophila Gld gene in seven species from the subgroup melanogaster. Dev Genet 15:38-50.

ROTHENBURG, S., F. KOCH-NOLTE, A. RICH, and F. HAAG. 2001. A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. Proc Natl Acad Sci USA 98:8985-8990.

RUEZ, C., F. PAYRE, and A. VINCENT. 1998. Transcriptional control of Drosophila bicoid by Serendipity delta: cooperative binding sites, promoter context, and co-evolution. Mech Dev 78:125-134.

SACCONE, G., I. PELUSO, D. ARTIACO, E. GIORDANO, D. BOPP, and L. C. POLITO. 1998. The Ceratitis capitata homologue of the Drosophila sex-determining gene Sex-lethal is structurally conserved, but not sex-specifically regulated. Development 125:1495-1500.

SACKERSON, C., M. FUJIOKA, and T. GOTO. 1999. The even-skipped locus is contained in a 16-kb chromatin domain. Dev Biol 211:39-52.

SAITO, T., H. M. LACHMAN, L. DIAZ et al. (and 9 co-authors). 2002. Analysis of monoamine oxidase A (MAOA) promoter polymorphism in Finnish male alcoholics. Psych Res 109:113-119.

SANDRELLI, F., S. CAMPESAN, M. G. ROSSETTO, C. BENNA, E. ZIEGER, A. MEGIGHIAN, M. COUCHMAN, C. P. KYRIACOU, and R. COSTA. 2001. Molecular dissection of the 5 ' region of no-on-transientA of Drosophila melanogaster reveals cis-regulation by adjacent dGpi1 sequences. Genetics 157:765-775.

SAUER, F., S. K. HANSEN, and R. TJIAN. 1995. DNA template requirement and activatorcoactivator requirements for transcriptional synergism by Drosophila bicoid. Science 270:1825-1827.

SCAFFIDI, P., and M. E. BIANCHI. 2001. Spatially precise DNA bending is an essential activity of the Sox2 transcription factor. J Biol Chem 276:47296-47302.

SCEMAMA, J. L., M. HUNTER, J. MCCALLAM, V. PRINCE, and E. STELLWAG. 2002. Evolutionary divergence of vertebrate Hoxb2 expression patterns and transcriptional regulatory loci. J Exp Zool 294:285-299.

SCHADT, E. E., S. A. MONKS, T. A. DRAKE et al. (and 11 co-authors). 2003. Genetics of gene expression surveyed in maize, mouse and man. Nature 422: 297-302.

SCHIFF, N. M., Y. FENG, J. A. QUINE, P. A. KRASNEY, and D. R. CAVENER. 1992. Evolution of the expression of the Gld gene in the reproductive tract of Drosophila. Mol Biol Evol 9:1029-1049.

SCHLICHTING, C. D., and M. PIGLIUCCI. 1998. Phenotypic evolution: A reaction norm perspective. Sinauer Associates, Inc., Sunderland, MA.

SEGAL, J. A., J. L. BARNETT, and D. L. CRAWFORD. 1999. Functional analysis of natural variation in Sp1 binding sites of a TATA-less promoter. J Mol Evol 49:736-749.

SEGIL, N., S. B. ROBERTS, and N. HEITZ. 1991. Mitotic phosphorylation of the Oct-1 homeodomain and regulation of Oct-1 DNA binding activity. Science 254:1814-1816.

SEOIGHE, C., N. FEDERSPIEL, T. JONES et al. (and 17 co-authors). 2000. Prevalence of small inversions in yeast gene order evolution. Proc Natl Acad Sci USA 97:14433-14437.

SERFLING, E., M. JASIN, and W. SCHAFFNER. 1985. Enhancers and eukaryotic genetranscription. Trends in Genetics 1:224-230.

SHABALINA, S. A., and A. S. KONDRASHOV. 1999. Pattern of selective constraint in C. elegans and C. briggsae genomes. Genet Res 74:23-30.

SHABALINA, S. A., A. Y. OGURTSOV, V. A. KONDRASHOV, and A. S. KONDRASHOV. 2001. Selective constraint in intergenic regions of human and mouse genomes. Trends Genet 17:373-376.

SHASHIKANT, C. S., C. B. KIM, M. A. BORBELY, W. C. WANG, and F. H. RUDDLE. 1998. Comparative studies on mammalian Hoxc8 early enhancer sequence reveal a baleen whale-specific deletion of a cis-acting element. Proc Natl Acad Sci USA 95:12364-12369.

SHAW, P. J., N. S. WRATTEN, A. P. MCGREGOR, and G. A. DOVER. 2002. Coevolution in bicoid-dependent promoters and the inception of regulatory incompatibilities among species of higher Diptera. Evol Dev 4:265-277.

SHIKAMA, N., J. LYON, and N. B. LA THANGUE. 1997. The p300/CBP family: Integrating signals with transcription factors and chromatin. Trends Cell Biol 7:230-236.

SHIN, H. D., C. WINKLER, J. C. STEPHENS et al. (and 12 co-authors). 2000. Genetic restriction of HIV-1 pathogenesis to AIDS by promoter alleles of IL10. Proc Natl Acad Sci USA 97:14467-14472.

SHORE, P., and A. D. SHARROCKS. 2001. Regulation of transcription by extracellular signals. Pp. 113-135 in J. Locker, ed. Transcription Factors. Academica Press, Inc., San Diego.

SIMON, J., M. PEIFER, W. BENDER, and M. O'CONNOR. 1990. Regulatory elements of the the bithorax complex that control expression along the anterior-posterior axis. Embo J 9:3945-3956.

SINGH, N., K. W. BARBOUR, and F. G. BERGER. 1998. Evolution of transcriptional regulatory elements within the promoter of a mammalian gene. Mol Biol Evol 15:312-325.

SINGH, N., and F. G. BERGER. 1998. Evolution of a mammalian promoter through changes in patterns of transcription factor binding. J Mol Evol 46:639-648.

SINHA, N. R., and E. A. KELLOGG. 1996. Parallelism and diversity in multiple origins of C4 photosynthesis in the grass family. Am J Bot 83:1458-1470.

SJOSTRAND, J. O., A. KEGEL, and S. U. ASTROM. 2002. Functional diversity of silencers in budding yeasts. Eukaryot Cell 1:548-557.

SJOTTEM, E., C. ANDERSEN, and T. JOHANSEN. 1997. Structural and functional analysis of DNA bending by Sp1 family transcription factors. J Mol Biol 267:490-504.

SKAER, N., and P. SIMPSON. 2000. Genetic analysis of bristle loss in hybrids between Drosophila melanogaster and D. simulans provides evidence for divergence of cisregulatory sequences in the achaete-scute gene complex. Dev Biol 221:148-167.

SKAER, N., D. PISTILLO, and P. SIMPSON. 2002. Transcriptional heterochrony of scute and changes in bristle pattern between two closely related species of blowfly. Dev Biol 252:31-45.

SMALE, S. T., A. JAIN, J. KAUFMANN, K. H. EMAMAI, K. LO, and I. P. GARRAWAY. 1998. The initiator element: a paradigm for core promoter heterogeneity within metazoan protein-coding genes. Cold Spring Harbor Symposium on Quatitative Biology 63:21-31.

SMALL, S., A. BLAIR, and M. LEVINE. 1992. Regulation of even-skipped stripe-2 in the Drosophila embryo. EMBO J 11:4047-4057.

SPEK, C. A., R. M. BERTINA, and P. H. REITSMA. 1999. Unique distance- and DNA-turndependent interactions in the human protein C gene promoter confer submaximal transcriptional activity. Biochem J 340:513-518.

STAUBER, M., H. JACKLE, and U. SCHMIDT-OTT. 1999. The anterior determinant bicoid of Drosophila is a derived Hox class 3 gene. Proc Natl Acad Sci USA 96:3786-3789.

STAUBER, M., A. PRELL, and U. SCHMIDT-OTT. 2002. A single Hox3 gene with composite bicoid and zerknult expression characteristics in non-Cyclorrhaphan flies. PNAS 99:274-279.

STERN, D. L. 2000. Perspective: Evolutionary developmental biology and the problem of variation. Evolution 54:1079-1091.

STERN, D. L. 1998. A role of Ultrabithorax in morphological difference between Drosophila species. Nature 396:463-466.

STOCKHAUS, J., U. SCHLUE, M. KOCZOR, J. A. CHITTY, W. C. TAYLOR, and P. WESTHOFF. 1997. The promoter of the gene encoding the C4 form of phoshpoenolpyruvate carboxylase directs mesophyll-specific expression in transgenic C4 Flaveria spp. Plant Cell 9:479-489.

STONE, J. R., and G. A. WRAY. 2001. Rapid evolution of cis-regulatory sequences via local point mutations. Mol Biol Evol 18:1764-1770.

STORGAARD, T., J. CHRISTENSEN, B. AASTED, and S. ALEXANDERSEN. 1993. Cis-acting sequences in the Aleutian mink disease parvovirus late promoter important for transcription: comparison to the canine parvovirus and minute virus of mice. J Virol 67:1887-1895.

STOUGAARD, J., N. N. SANDAL, A. GRØN, A. KUHLE, and K. A. MARCKER. 1987. 5' analysis of the soybean leghaemoglobin lbc3 gene: Regulatory elements required for promoter activity and organ specificity. EMBO J 6:3565-3569.

STREELMAN, J. T., and T. D. KOCHER. 2002. Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia. Physiol Genomics 9:1-4.

STRUHL, K. 1999. Fundamentally different logic of gene regulation in eukaryotes and

prokaryotes. Cell 98:1-4.

SUCENA, E., and D. L. STERN. 2000. Divergence of larval morphology between Drosophila sechellia and its sibling species caused by cis-regulatory evolution of ovo/shaven-baby. Proc Natl Acad Sci USA 97:4530-4534.

SUSKE, G. 1999. The Sp-family of transcription factors. Gene 238:291-300.

SUTTON, K. A., and M. WILKINSON. 1997. Rapid evolution of a homeodomain: Evidence for positive selection. J Mol Evol 45:579-588.

SWALLA, B. J., and W. R. JEFFERY. 1996. Requirement of the Manx gene for expression of chordate features in a tailless ascidian larva. Science 274:1205-1208.

TAJIMA, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.

TAKAHASHI, A., S. C. TSAUR, J. A. COYNE, and C. I. WU. 2001. The nucleotide changes governing cuticular hydrocarbon variation and their evolution in Drosophila melanogaster. Proc Natl Acad Sci USA 98:3920-3925.

TAKAHASHI, H., Y. MITANI, G. SATOH, and N. SATOH. 1999. Evolutionary alterations of the minimal promoter for notochord-specific Brachyury expression in ascidian embryos. Development 126:3725-3734.

TAKAHATA, N. 1987. On the overdispersed molecular clock. Genetics 116:169-179.

TAMARINA, N. A., M. Z. LUDWIG, and R. C. RICHMOND. 1997. Divergent and conserved features in the spatial expression of the Drosophila pseudoobscura esterase-5B gene and the esterase-6 gene of Drosophila melanogaster. Proc Natl Acad Sci USA 94:7735-7741.

TAUTZ, D. 2000. Evolution of transcriptional regulation. Curr Op Genet Dev 10:575-579.

THANOS, D., and T. MANIATIS. 1995. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. Cell 83:1091-1100.

THEISSEN, G., A. BECKER, A. DI ROSA, A. KANNO, J. T. KIM, T. MUNSTER, K. U. WINTER, and H. SAEDLER. 2000. A short history of MADS-box genes in plants. Plant Mol Biol 42:115-149.

THOMPSON, J. R., S. W. CHEN, L. HO, A. W. LANGSTON, and L. J. GUDAS. 1998. An evolutionary conserved element is essential for somite and adjacent mesenchymal expression of the Hoxa1 gene. Dev Dyn 211:97-108.

THURZ, M. 2001. Genetic susceptibility in chronic viral hepatitis. Antiviral Res 52:113-116.

TING, C. T., S. C. TSAUR, M. L. WU, and C. I. WU. 1998. A rapidly evolving homeobox at the site of a hybrid sterility gene. Science 282:1501-1504.

TOMAREV, S. I., M. K. DUNCAN, H. J. ROTH, A. CVEKL, and J. PIATAGORSKY. 1994. Convergent evolution of crystallin gene regulation in squid and chicken: the AP-1/ARE connection. J Mol Evol 39:134-143.

TORCHIA, J., C. K. GLASS, and M. G. ROSNFELD. 1998. Co-activators and co-repressors in the integration of transcriptional responses. Curr Opin Cell Biol 10:373-383.

TOURNAMILLE, C., Y. COLIN, J. P. CARTRON, and C. LE VAN KIM. 1995. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. Nat Genet 10:224-228.

TREFILOV, A., J. BERARD, M. KRAWCZAK, and J. SCHMIDTKE. 2000. Natal dispersal in rhesus macaques is related to serotonin transporter gene promoter variation. Behav Genet 30:295-301.

TREISMAN, J., P. GÖNCZY, M. VASHISHTHA, E. HARRIS, and C. DESPLAN. 1989. A single amino acid can determine the DNA binding specificity of homeodomain proteins. Cell 59:553-562.

TRIEZENBERG, S. J. 1995. Structure and function of transcriptional activation domains.

Curr Op Gen Dev 5:190-196.

TÜMPEL, S., M. MACONOCHIE, L. M. WIEDEMANN, and R. KRUMLAUF. 2002. Conservation and diversity in the cis-regulatory networks that integrate information controlling expression of Hoxa2 in hindbrain and cranial neural crest cells in vertebrates. Dev Biol 246:45-56.

VARGA-WEISZ, P. 2001. ATP-dependent chromatin remodeling factors: nucleosome shufflers with many missions. Oncogene 20:3076-3085.

VENTER, J. C.M. D. ADAMSE. W. MYERS et al. (and 271 co-authors). 2001. The sequence of the human genome. Science 291:1304-1351.

VIDIGAL, P., J. J. GEMNER, and N. N. ZEIN. 2002. Polymorphsisms in the interleukin-10, tumor necrosis factor-a, and transforming growth factor-b1 genes in chronic hepatitis C patients treated with interferon and ribavarin. J Hepatol 36:271-277.

VOGELAUER, M., J. WU, N. SUKA, and M. GRUNSTEIN. 2000. Global histone acetylation and deacetylation in yeast. Nature 408:495-498.

VON DASSOW, G., E. MEIR, E. M. MUNRO, and G. M. ODELL. 2000. The segment polarity network is a robust developmental module. Nature 406:188-192.

WAGNER, A. 2001. The yeast protein interaction network evolves rapidly and contains few duplicate genes. Mol Biol Evol 18:1283-1292.

WALTER, J., and M. D. BIGGIN. 1996. DNA binding specificity of two homeodomain proteins in vitro and in Drosophila embryos. Proc Natl Acad Sci USA 93:2680-2685.

WANG, R. L., A. STEC, J. HEY, L. LUKENS, and J. DOEBLEY. 1999. The limits of selection during maize domestication. Nature 398:236-239.

WANG, W., F. G. BRUNET, E. NEVO, and M. LONG. 2002. Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. PNAS 99:4448-4453.

WATERSTON, R. H.K. LINDBLAD-TOHE. BIRNEY et al. (and 219 co-authors). 2002. Initial

sequencing and comparative analysis of the mouse genome. Nature 420:520-562.

WEI, Z., L. M. ANGERER, M. L. GAGNON, and R. C. ANGERER. 1995. Characterization of the SpHE promoter that is spatially regulated along the animal-vegetal axis of the sea urchin embryo. Dev Biol 171:195-211.

WEINZIERL, R. O. J. 1999. Mechanisms of gene expression. Imperial College Press, London.

WEIS, L., and D. REINBERG. 1992. Transcription by RNA polymerase II: Initiator-directed formation of transcription-competent complexes. FASEB J 6:3300-3309.

WEST, R. J., R. YOCUM, and M. PTASHNE. 1984. Saccharomyces cerevisiae GAL1-GAL10 divergent promoter region: location and fuction of the upstream activating sequence UASG. Mol Cell Biol 4: 2467-2478.

WHEELER, J. C., K. SHIGESADA, J. P. GERGEN, and Y. ITO. 2000. Mechanisms of transcriptional regulation by Runt domain proteins. Semin Cell Dev Biol 11:369-375.

WHITE, K. P., S. A. RIFKIN, P. HURBAN, and D. S. HOGNESS. 1999. Microarray analysis of Drosophila development during metamorphosis. Science 286:2179-2184.

WHITE, R. J. 2001. Gene transcription: Mechanisms and control. Blackwell Science, Malden, MA.

WILKINS, A. S. 1993. Genetic analysis of animal development. Wiley-Liss, Inc., New York.

WILKINS, A. S. 2002. The evolution of developmental pathways. Sinauer Associates, Sunderland, MA.

WILSON, A. C. 1975. Evolutionary importance of gene regulation. Stadler Symp 7:117-134.

WILTON, A. N., C. C. LAURIE-AHLBERG, T. H. EMIGH, and J. W. CURTSINGER. 1982. Naturally occurring enzyme activity variation in Drosophila melanogaster. II.

Relationships among enzymes. Genetics 102:207-221.

WOLFF, C., M. PEPLING, P. GERGEN, and M. KLINGLER. 1999. Structure and evolution of a pair-rule interaction element: runt regulatory sequences in D. melanogaster and D. virilis. Mech Dev 80:87-99.

WOLFFE, A. P. 2001. Chromatin structure and the regulation of transcription. Pp. 35-64 in J. Locker, ed. Transcription Factors. Academic Press, Inc., San Diego.

WOLFFE, A. P. 1994. Insulating Chromatin. Curr Biol 4:85-87.

WRAY, G. A., and A. E. BELY. 1994. The evolution of echinoderm development is driven by several distinct factors. Development 120, Supplement:97-106.

WRAY, G. A., and C. J. LOWE. 2000. Developmental regulatory genes and echinoderm evolution. Sys Biol 49:28-51.

WRAY, G. A., and D. R. MCCLAY. 1989. Molecular heterochronies and heterotopies in early echinoid development. Evolution 43:803-813.

WRIGHT, C. E., F. HADDAD, A. X. QUIN, P. W. BODELL, and K. M. BALDWIN. 1999. In vivo regulation of ß-MHC gene in rodent heart: Role of T3 and evidence for an upstream enhancer. Am J Physiol 276:C883-C891.

WRIGHT, S. 1982. Character change, speciation, and the higher taxa. Evolution 36:427-443.

WU, C. Y., and M. D. BRENNAN. 1993. Similar tissue-specific expression of the Adh genes from different Drosophila species is mediated by distinct arrangements of cis-acting sequences. Mol Gen Genet 240:58-64.

XU, P.-X., X. ZHANG, S. HEANEY, A. YOON, A. M. MICHELSON, and R. L. MAAS. 1999. Regulation of Pax6 expression is conserved between mice and flies. Development 126:383-395.

YAMAMOTO, K. R., B. D. DARIMONT, R. L. WAGNER, and J. A. IÑIGUEZ-LLUHÍ. 1998. Building transcriptional regulatory complexes: Signals and surfaces. Cold Spring Harbor Symposium on Quatitative Biology 63:587-598.

YAMAMOTO, Y., and W. R. JEFFERY. 2000. Central role for the lens in cave fish eye degeneration. Science 289:631-633.

YU, H., S. H. YANG, and C. J. GOH. 2002. Spatial and temporal expression of the orchid floral homeotic gene DOMADS1 is mediated by its upstream regulatory regions. Plant Mol Biol 49:225-237.

YUH, C.-H., C. T. BROWN, C. B. LIVI, L. ROWEN, P. J. C. CLARKE, and E. H. DAVIDSON. 2002. Patchy interspecific sequence similiarities efficiently identify positive cis-regulatory elements in the sea urchin. Dev Biol 246:148-161.

YUH, C. H., H. BOLOURI, and E. H. DAVIDSON. 1998. Genomic cis-regulatory logic: Experimental and computational analysis of a sea urchin gene. Science 279:1896-1902.

YUH, C. H., H. BOLOURI, and E. H. DAVIDSON. 2001. Cis-regulatory logic in the endo16 gene: Switching from a specification to a differentiation mode of control. Development 128:617-629.

YUN, K. S., and B. WOLD. 1996. Skeletal muscle determination and differentiation: Story of a core regulatory network and its context. Curr Op Cell Biol 8:877-889.

ZELLER, R. W., J. D. GRIFFITH, J. G. MOORE, C. V. KIRCHHAMER, R. J. BRITTEN, and E. H. DAVIDSON. 1995. A multimerizing transcription factor of sea urchin embryos capable of looping DNA. Proc Natl Acad Sci USA 92:2989-2993.

ZERUCHA, T., T. STUHMER, G. HATCH et al. (and 7 co-authors). 2000. A highly conserved enhancer in the Dlx5/Dlx6 region is the site of cross-regulatory interactions between Dlx genes in the embryonic forebrain. J Neuro 20:709-721.

ZUCKERKANDL, E. 1963. Perspectives in molecular anthropology. Pp. 256-274 in A. Rich, and N. Davidson, eds. Structural chemistry and molecular biology. W. H. Freeman, San Francisco.

Zwarts et al. (2002) ABCA1 regulatory variants influence coronary artery disease independent of effects on plasma lipid levels. Clin. Genet. 61:115-125

The present invention is not to be limited in scope by the specific embodiments described herein. Indeed, various modifications of the invention in addition to those described herein will become apparent to those skilled in the art from the foregoing description and accompanying figures. Such modifications are intended to fall within the scope of the appended claims.

Various references, including patent applications, patents and publications are cited herein, the disclosures of which are incorporated by reference in their entireties.

CLAIMS

1. A method for stratifying a patient in a subgroup of a clinical trial for the prevention or treatment of a disease selected from the diseases listed in column 12 of Table 1, comprising determining the genotype of a functional site listed for said disease in column 12 of Table 1, and stratifying the patient in a subgroup of a clinical trial according to said genotype.

2. An isolated polynucleotide comprising a sequence selected from the group consisting of:
   (a) a sequence provided in Table 1;
   (b) a complement of a sequence provided in Table 1;
   (c) a sequence consisting of at least 10 contiguous residues of a sequence provided in Table 1;
   (d) a sequence that hybridizes to a sequence provided in Table 1 or its complement, under moderately stringent conditions; and
   (e) a sequence having at least 75% identity to a sequence of Table 1;
   wherein said sequence is not flanked at its 5' end or its 3' end in said polynucleotide by greater than 200 nucleotides of sequences that are contiguous to said sequence in the human genome.

3. The polynucleotide of claim 2, wherein said sequence is not flanked at its 5' end or its 3' end by greater than 100 nucleotides of sequences that are contiguous to said sequence in the human genome.

4. The polynucleotide of claim 3, wherein said sequence is not flanked at its 5' end or its 3' end by greater than 50 nucleotides of sequences that are contiguous to said sequence in the human genome.

5. The polynucleotide of claim 4, wherein said sequence is not flanked at its 5' end or its 3' end by greater than 10 nucleotides of sequences that are contiguous to said sequence in the human genome.

6. A vector comprising a polynucleotide of any one of claims 2-5.

7. The vector of claim 6, wherein the polynucleotide is operably linked to an open reading frame.

8. The vector of claim 7, wherein the open reading frame encodes a reporter.

9. The vector of claim 8, wherein the reporter is selected from the group consisting of: alkaline phosphatase, β-galactosidase, neomycin phosphotransferase, chloramphenicol acetyltransferase, dihydrofolate reductase, hygromycin phosphotransferase, beta-glucoronidase, green fluorescent protein, and luciferase genes.

10. A host cell transformed or transfected with a vector according to claim 2.

11. A host cell transformed or transfected with a ploynucleotide according to claim 2.

12. An isolated polynucleotide comprising a plurality of polynucleotides of any one of claims 2-5.

13. A non-human animal comprising a vector of claim 6.

14. The animal of claim 13, wherein said animal is a mammal.

15. A cell or non-human organism in which a coding sequence that is natively operably linked to a sequence provided in Table 1 is replaced by a different coding sequence.

Figure 1.

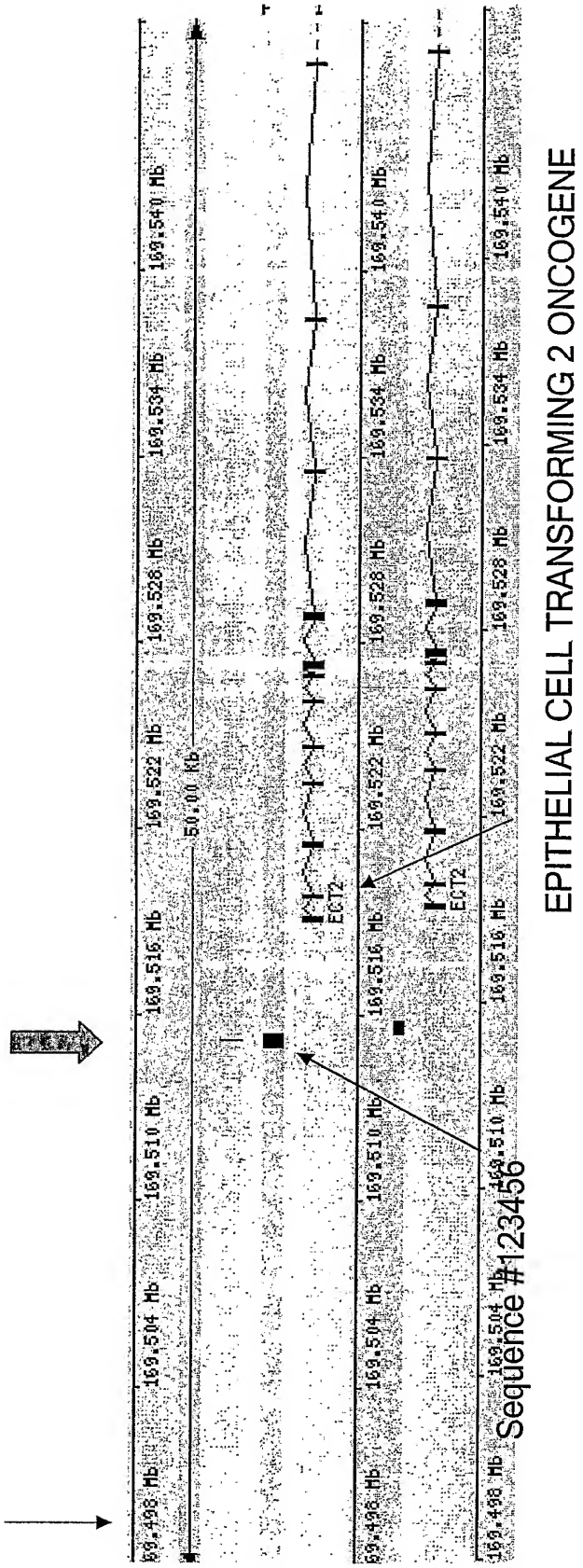

Position of confirmed hypersensitive site
Chromosome 3 position
169.498 Mb
169.504 Mb
169.510 Mb
169.516 Mb
169.522 Mb
169.528 Mb
169.534 Mb
169.540 Mb
50.00 Kb
ECT2
ECT2
Sequence #123456
EPITHELIAL CELL TRANSFORMING 2 ONCOGENE

Figure 2.

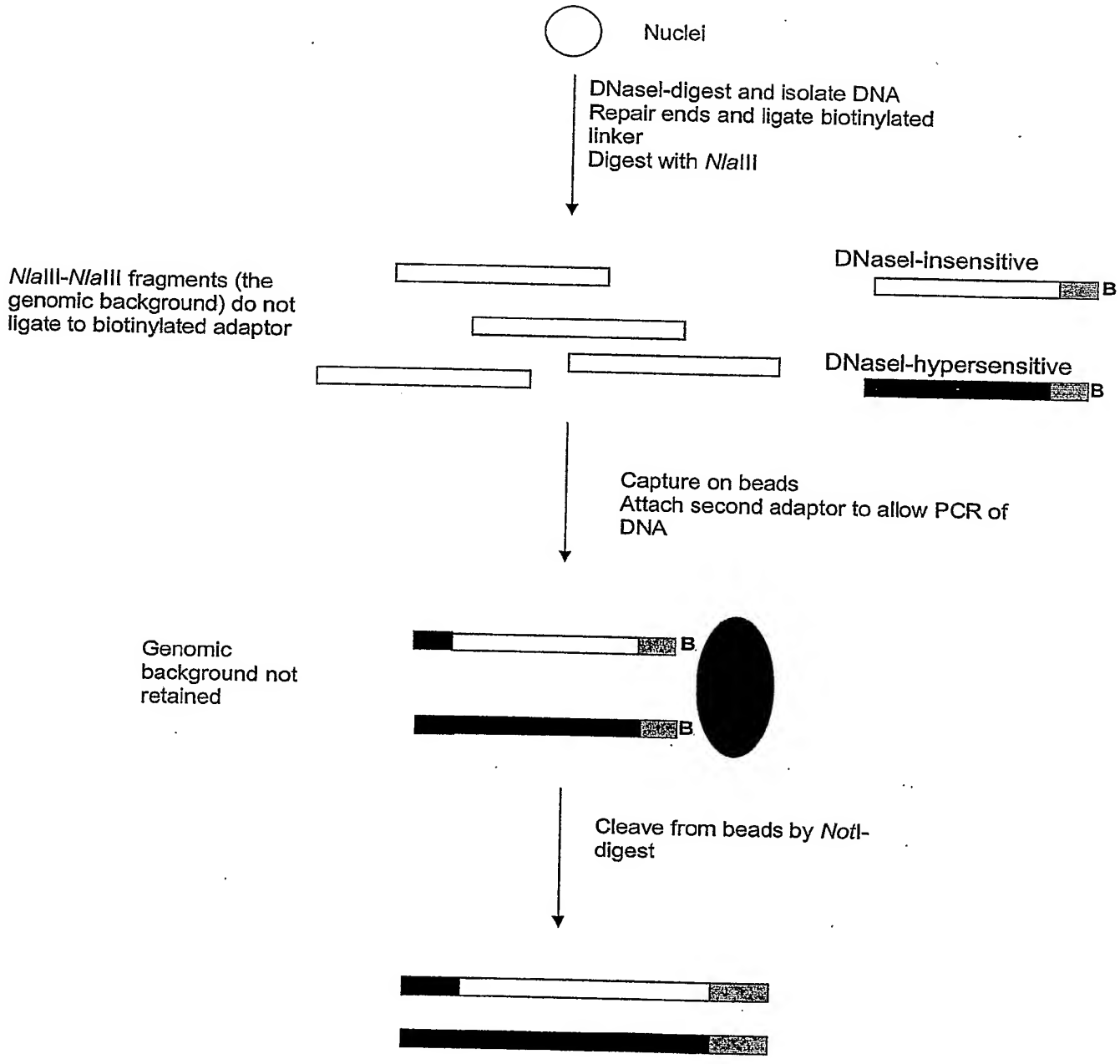
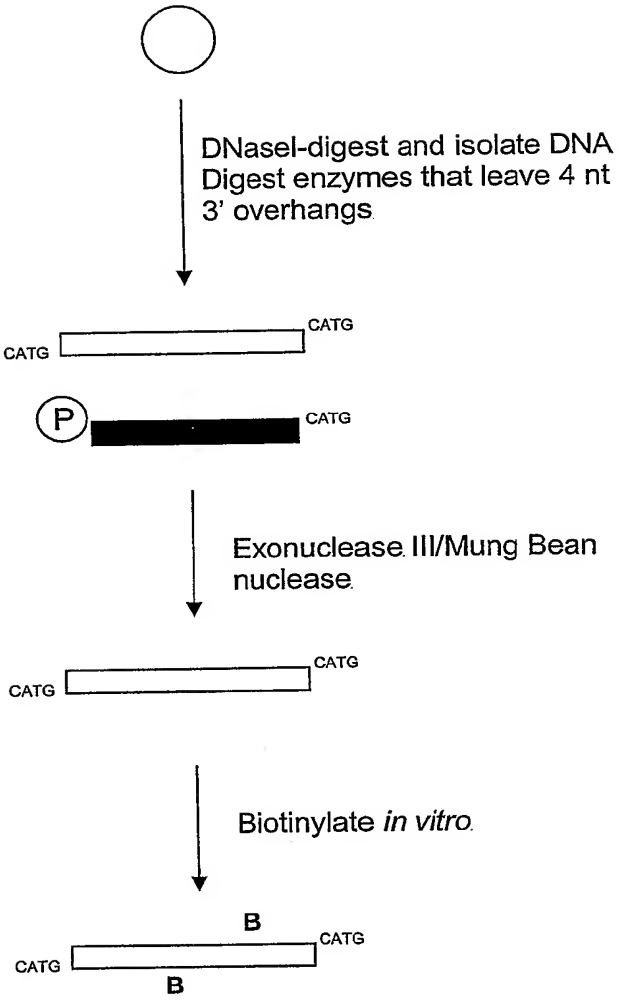

Nuclei
DNaseI-digest and isolate DNA
Repair ends and ligate biotinylated linker
Digest with NlaIII
NlaIII-NlaIII fragments (the genomic background) do not ligate to biotinylated adaptor
DNaseI-insensitive
B
DNaseI-hypersensitive
B
Capture on beads
Attach second adaptor to allow PCR of DNA
Genomic background not retained
B
B
Cleave from beads by NotI-digest

Figure 3.


DNaseI-digest and isolate DNA
Digest enzymes that leave 4 nt
3' overhangs
CATG
CATG
P
CATG
Exonuclease III/Mung Bean
nuclease
CATG
CATG
Biotinylate *in vitro*
B
CATG
CATG
B